The role of non-genetic variability in Acute Myeloid Leukaemia

Shikha Gupta

Department of Genetics
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Murray Edwards College February 2021

I dedicate this work to my loving parents, Saroj Gupta, Babu Lal Gupta and my brother Ankur Gupta.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared and specified in the text.

It does not exceed the prescribed word limit for the Biology Degree Committee.

Shikha Gupta

February 2021

Abstract

The role of non-genetic variability in Acute Myeloid Leukaemia

Acute myeloid leukaemia (AML) is a heterogeneous clonal disorder of haematopoietic progenitor cells with a dismal survival. It has a strong reliance on epigenetic and transcriptional factors for disease progression. Accordingly, my lab has previously identified *KAT2A*, a histone acetyl-transferase, as a requirement for AML maintenance, where chemical inhibition of KAT2A promotes differentiation of AML cell lines (Tzelepis *et al.*, 2016). More recently, using a conditional knockout mouse model for *Kat2a* our lab showed that it sustains *KMT2A/MLLT3* AML stem cells. *Kat2a* is a classical regulator of transcriptional variability, its loss leading to cell-to-cell heterogeneity in transcription levels, including from genes involved in ribosomal biogenesis and translation (Domingues *et al.*, 2020). No recurrent mutations in the *KAT2A* gene have been described in AML, and it is unclear if and how it participates in pre-leukaemia-to-AML progression. In this thesis, I studied *Kat2a* loss in 2 mouse models of AML representing forms of human disease with a prolonged pre-leukaemia phase which typically require additional mutations for leukaemia progression. Specifically, I analysed the biology of *RUNX1-RUNX1T1(9a)* and *Idh1R132H*-initiated AML in a conditional *Kat2a*KO background and observed consistent acceleration of leukaemia initiation and progression with perpetuation of transformed *Kat2a*KO cells *in vivo*. Single-cell RNA sequencing (scRNA-seq) of early-stage *Kat2a*WT and *Kat2a*KO *RUNX1-RUNX1T1(9a)* pre-leukaemia, suggested an increase in transcriptional variability upon *Kat2a* loss, which was accompanied by diversification of cell fates towards B-lymphocytes and monocytes. Furthermore, pseudo-temporal ordering of single *Kat2a*KO cells revealed a highly branched trajectory populated with intermediate stages of transformation, including accumulation of leukaemia progenitors with *RUNX1-RUNX1T1* signature. In contrast, *Kat2a*WT cells displayed a linear haematopoiesis trajectory with minimal branching, and an abrupt transition towards the candidate leukaemia progenitor state. Pathway analysis combined with functional studies indicate a mechanistic contribution of cytoplasmic translation and ribosomal biogenesis-associated genes towards leukaemia progression in both models of pre-leukaemia. Taken together, my work suggests that loss of *Kat2a* results in accelerated pre-leukaemia transformation accompanied with diversification of cell fate transitions including with increased accessibility to cell states prone to transformation. Furthermore, transformation-prone cells may benefit from low biosynthetic activity to progress to a leukaemic state. I hypothesize that *Kat2a* loss may function similarly in the context of other malignancies. In the future, this knowledge may aid in the development of early diagnostic tools and suggest bespoke therapeutic interventions.

Shikha Gupta

Acknowledgements

I feel humbled and grateful to acknowledge all those who have helped me to frame ideas concretely. I am grateful to my supervisor Dr. Cristina Pina for providing me the opportunity to work on this study and would like to thank her for constructive comments during thesis writing. I express my warmest gratitude to Dr. Gos Micklem for accepting me as his nominal student since August 2019 and am indebted to him for his time and efforts in ensuring my well-being and for providing constructive suggestions whenever needed. I appreciate the help and support received from my previous advisor Dr. Sudhakaran Prabakaran during the first year of my PhD and would remain thankful to him for introducing me to the basics of next-generation sequencing technology and terminologies.

I would like to thank our collaborator Prof. George Vassiliou (WT Sanger Institute) for sharing *Idh1*R132H mouse model for this study and his lab member Dr. Oliver Dovey for developing the model. I feel delighted to have worked with Oliwia Cyran, placement student from King's College London who worked towards the *in vitro* pre-leukaemia studies of *Idh1*R132H mouse model with assistance from Ryan J. Asby (Huntly group, Department of Haematology). I am also thankful to Caitlin Cash, student at Brunel University, London, for working towards leukaemia analysis of *Idh1*R132H mouse model while I've been working from my home country during the pandemic. I would like to thank Ana Filipa Domingues (ex-member Pina group) and George Giotopolous (Huntly group) for working on the preliminary set-up of the *RUNX1-RUNX1T1(9a)* mouse model before I joined the group. I would like to extend my gratitude towards Dr. Roberto Bandiera (Frye group) for sharing reagents for the OP-Puro experiment and providing his help in experimental set up. I would further like to thank Dr. Shabana Vohra (Huntly group) for running some pre-processing steps on single cell ATAC sequencing data before my lab joining and to Dr. Matt Wayland (Department of Zoology) for providing high performance computing resources to run the CellRanger pipeline for single-cell RNA sequencing data. I am grateful to Joana Ceveira (Flow cytometry manager, Department of Pathology) for being extremely helpful and teaching me how to operate the Attune flow analyzer in a short period of time.

I would like to express my sincere gratitude to my previous supervisor (2013-2015), Dr. Sweta Srivastava (Assistant Professor, St. John's Medical College Hospital, India) for her guidance and motivation, which encouraged me to follow a scientific career. I am thankful to her for inculcating in me the importance of laboratory management, experimental planning, multi-tasking, and teamwork skills which proved vital in finishing my thesis work. In particular, I am grateful for the education I received from her which helped me navigate challenging circumstances in a new country, which included a department and supervisor change, funding constraints in the lab, and the onset of a pandemic

during the stage of thesis conclusion. I will remain indebted to her for helping me become an independent researcher and I feel lucky to have the chance to be associated with someone like her in my life.

I express my heartfelt gratitude to my Graduate Tutor, Dr. Emma Cahill (Murray Edwards College) for her constant encouragement and support, especially during challenging times. I can't thank her enough for her insightfulness and empathy which helped me in finishing my thesis. I would like to extend my gratitude towards my General Practitioner, Dr. Rebecca Towl and Ms. Esther at University Counselling Services for their excellent care and lending an ear when in need.

I owe my gratitude to my partner Easwaran Ramamurthy (PhD student, Carnegie Mellon University, USA). His constant encouragement helped me keep pursuing my interest in Science. Considering the lack of computational guidance and environment in my lab, his availability despite the time zone difference to answer my questions helped me pick up R, Python, and UNIX which I used for analysing sequencing data in this thesis. His care and support during this time have helped me get through difficult times.

I could not have asked for better lab members than Keti, Liliana, and Wade who went through hard times together, cheered me on, and celebrated each accomplishment. Their timely help and friendship will always be remembered. I would also like to thank members from neighbouring labs including Chapman group (Haematology), Ferguson Smith group and Imbeault group (Genetics) for their help whenever needed.

I am extremely thankful to my friends Shilpa, Akshit, Ying, Chilombo, Amit, Pallavi, Radhika, Avinash, Grace, Shalini, and Elizabeth for their love, care and compassionate friendship. I will always cherish the warmth shown by them. I am also grateful to Sarah Horton (Huntly group), Rachel Lyne (Micklem group) and Susanne van den Brink (van Oudenaarden group, Hubrecht Institute) for their comments and suggestions which helped me improve my thesis.

This work wouldn't have been possible without the prompt support received from the administrative staff at Department of Haematology, Department of Genetics, and Murray Edwards College. I would especially like to thank Martin Dawes for his truly professional support.

I feel a deep sense of gratitude for my parents who formed part of my vision and taught me good values and about the things that matter in life. Their infallible love and support have always been my strength. Their patience and sacrifice will remain an inspiration for me throughout my life. I am extremely

grateful to my brother Ankur for his unconditional trust, timely encouragement, and support despite the long distance between us, and to my niece Anvika for her cheerfulness.

I acknowledge Lady Tata Memorial Trust, Trinity Henry Barlow Trust, Cambridge Commonwealth, European and International Trust and Murray Edwards College for believing in my work and providing financial support during the course of my PhD without which it would have been impossible to pursue my degree.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AF	Alexa Fluor
AF9	ALL1-Fused gene from chromosome 9
ALL	Acute Lymphoblastic Leukaemia
AML	Acute Myeloid Leukaemia
AML1	Acute Myeloid Leukaemia 1
ASPA	Animal Scientific Procedures Act 1986
ATAC	Ada Two A Containing
AWERB	Animal Welfare and Ethical Review Body
BCL	Base Call Files
BE	Branching Evolution
BED	Browser Extensible Data
BM	Bone Marrow
BRD	Bromodomain
BSA	Bovine Serum Albumin
CDS	Cell Data Set
cDNA	complimentary DNA
CE	Convergent Evolution
CEL-seq	Cell Expression by Linear Amplification Sequencing
CFC	Colony Forming assay
CFP	Cyan Fluorescent Protein
CFU-E	ColonyForming Unit-erythroid
CFU-G	Colony Forming Unit-Granulocyte
CFU- GEMM	Colony Forming Unit- Granulocyte Erythroid Macrophage Megakaryocyte
CFU-GM	Colony Forming Unit-Granulocyte Macrophage
CLL	Chronic Lymphoid Leukaemia
CO_2	Carbon Dioxide
CNV	Copy-Number Variation
CTC	Circulating Tumour Cell
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
D10	DMEM + 10% FBS+ 2 mg/mL L-Glutamine, 1% PSA
D-2-HG	D-2-Hydroxyglutarate

DBA	Diamond-Blackfan anaemia
DLBCL	Diffuse Large B-cell Lymphoma
DMEM	Dulbecco's Modified Eagle's medium
DM	Distance to Median
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
DUB	Histone Deubiquitinase
EDTA	Ethylenediaminetetraacetic Acid
ETO	Eight Twenty One
FACS	Fluorescence Activated Cell Sorting
FBS	Fetal Bovine Serum
FDR	False Discovery Rate
FLT3	Fms-Related Tyrosine Kinase 3
Gcn5	General control nonderepressible 5
GFP	Green Fluorescent Protein
GREAT	Genomic Regions Enrichment of Annotations Tool
GNAT	Gcn5-related N-acetyltransferase
G&T-seq	Genome and Transcriptome sequencing
HAT	Histone Acetyl Transferase
HDAC	Histone Deacetylase
HEK	Human Embryonic Kidney
HGB	Haemoglobin
HIF1α	Hypoxia inducible factor 1 α
I20	IMDM + 20% FBS + 2 mg/mL L-Glutamine, 1% PSA
IMDM	Iscove's Modified Dulbecco's Medium
IRES	Internal Ribosome Entry Site
iPSCs	induced Pluripotent Stem Cells
IVT	*in vitro* transcription
Kat2a	Lysine Acetyltransferase 2a
KO/NULL	Knockout
LE	Linear Evolution
log_2 FC	log_2 fold change
LSC	Leukaemia Stem Cell

LT-HSC	Long-Term reconstituting Haematopoietic Stem Cell
M	Macrophage
M-MLV	Moloney murine leukaemia virus
MALBAC	Multiple Annealing and Looping-Based Amplification Cycles
MARS-seq	Massively Parallel RNA Single-cell sequencing
MB-3	α-methylene-γ-butyrolactone 3
mIL-3	mouse Interleukin-3
mIL-6	mouse Interleukin-6
MLL	Mixed Lineage Leukaemia
MMHCC	Mouse Model of Human Cancers Consortium
MOMP	Mitochondrial Outer Membrane Permeabilization
mRNA	messenger RNA
mSCF	mouse-Stem Cell Factor
MSigDB	Molecular Signature Database
NaCl	Sodium Chloride
NAD^+	Nicotinamide adenine dinucleotide
$NADP^+$	Nicotinamide Adenine Dinucleotide Phosphate
NAT	N-terminal and Ada-Two interaction domain
NE	Neutral Evolution
NGS	Next Generation Sequencing
NHR	Nervy homology regions
NOD/SCID	Nonobese diabetic/severe combined immunodeficiency
NPM1	Nucleophosmin 1
NRAS	Neuroblastoma RAS Viral (V-Ras) Oncogene Homolog
OP-Puro	O-propargyl-puromycin
OXPHOS	Oxidative Phosphorylation
PAGA	Partition-based Graph Abstraction
PANTHER	Protein Analysis through Evolutionary Relationships
PBE	Phosphate Buffer Saline + 0.5M EDTA + 2% BSA
PBS	Phosphate Buffer Saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PDR	Proportion Discordant Reads

PE	Punctuated Evolution
PFA	Paraformaldehyde
pIpC	Polyinosinic polycytidylic acid
PKA	Protein Kinase A
PKA RIIalpha	PKA regulatory subunit
PLT	Platelets
PSA	Penicillin-Streptomycin Antibiotic
PCAF	p300/CBP-associated factor
QC	Quality Control
qRT-PCR	Quantitative Reverse Transcription Polymerase Chain Reaction
R20	RPMI + 20% FBS
RBC	Red Blood Cell
RNA	Ribonucleic acid
Rpm	Rotations per minute
RPMI 1640	Roswell Park Memorial Institute 1640
rRNA	ribosomal RNA
S6K1	p70 ribosomal S6 kinase
scATAC-seq	Single cell sequencing Assay for Transposase-Accessible Chromatin sequencing
scM&T-seq	Single cell genome-wide Methylome and Transcriptome sequencing
scRNA-seq	Single cell RNA sequencing
scTrio-seq	Single cell Triple Omics sequencing
SDS	Sodium dodecyl sulphate
SMART	Switching Mechanism At 5′ end of RNA Template
SNV	Single Nucleotide Variation
SPF	Specific Pathogen Free
SAGA	Spt-Ada-Gcn5 acetyltransferase
T-ALL	T-cell acute lymphoblastic leukaemia
TAF	TATA-binding protein associated factor
TCA	Tricarboxylic Acid
TCGA	The Cancer Genome Atlas
TET	Ten-eleven translocation
TMRE	Tetramethylrhodamine ethyl ester

Tris-Cl	Tris(hydroxymethyl)aminomethane hydrochloride
TRRAP	Transformation/Transcription domain-Associated Protein
tSNE	t-distributed stochastic neighbour embedding
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecule Identifier
WBC	White Blood Cell
WGS	Whole-genome sequencing
WT	Wild type
YFP	Yellow fluorescent protein
α-KG	α-Ketoglutarate
2-HG	2-Hydroxyglutarate

1 Introduction

1.1 Cancer Evolution

Cancer is a disease with dysregulated growth and survival. It is the second leading cause of death globally with 1 in 6 deaths occurring due to cancer (Roth *et al.*, 2018). There were 17 million cases of cancer diagnosed around the world, including 9.5 million deaths, in the year 2018 and the global burden of new cancer cases is expected to reach 27.5 million including 16.2 million cancer deaths by 2040 (Bray *et al.*, 2018). About 20% of the cancer cases are found to be in low- and medium Human Development Index Countries which have scarcity of medical resources and lack of supportive health system to circumvent the disease burden. One of the major reasons for this high mortality rate is the high failure rates in oncological drug discovery for targeted and personalized medicine development. Despite the advent of high-throughput tumour analytical techniques, biomarker validation and clinical qualification of such biomarkers remains challenging. Unfortunately, less than 1% of all reported cancer biomarkers enter clinical practice (Kern, 2012). The challenges posed by difficulties in biomarker research can be approached by studying the cancer disease progression from an evolutionary perspective.

In this context, the theory of natural selection laid by Charles Darwin suggesting “the fittest will survive, and a race will be eventually produced adapted to the conditions in which it lives” (Gerlinger, Marco,Nicholas Mcgranahan, 2014) is a paramount in understanding the concept of evolution in more detail. The fundamental principles of Darwinian evolution were originally framed in relation to the evolution of unicellular or multicellular organisms within a given population until Peter C. Nowell explained that the evolutionary concept is equally valid in the context of cancer evolution (Nowell, 1976). He suggested that acquired genetic lability permits stepwise selection of variant sublines and underlies tumour progression (Fig 1.1). His hypothesis that natural selection occurs in cancer in the form of clonal selection which develops from a single cell of origin leading to constant evolutionary change and possibly drug resistance, gained a widespread audience. Evolutionary studies of clonal cancer development have since been conducted, allowing remarkable biological inferences to be made (Gerlinger and Swanton, 2010; Greaves and Maley, 2012; Yates and Campbell, 2012).

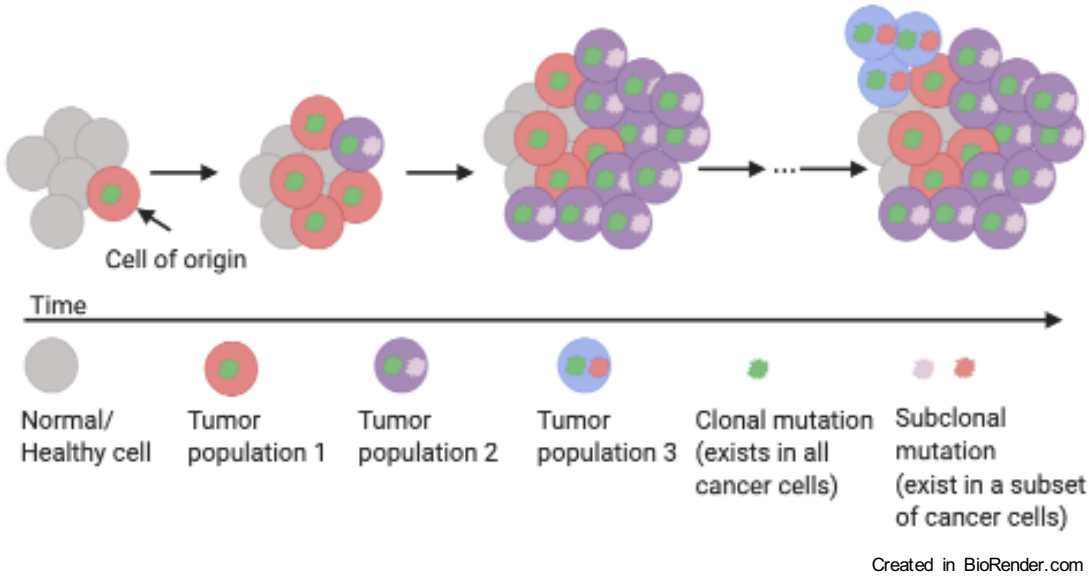


Figure 1.1: Schematic representing theory of clonal evolution of cancer led by Nowell (Nowell, 1976).

Clonal evolution of cancer is primarily shaped by the following fundamental forces-

1. Mutation
2. Genetic drift
3. Selection

Mutation is a stochastic process by which a change in genetic code is introduced into the given population of cells. Genetic drift is a consequence of mutational events and describes the stochastic changes in clone size due to random effects that lead to cancer cell growth or death. Selection is deterministic and is the fixation of a mutation and its associated clonal characteristics, which happen if they result in adaptive advantage (Szendro *et al.*, 2013).

Mutation is crucial for the evolution of cancer, as the diversity generated as a result of somatic genetic aberrations in cancer cells enables selection. These mutations may vary from point mutations to insertion/ deletion of several base pairs or rearrangements of entire chromosomal segments and can modify the protein-coding regions modifying the resultant protein with altered function or making it non-functional. Based on their contribution to cancer cell fitness, these mutations can be defined as driver or passenger aberrations. Driver mutations, as the name suggests, confer a selective advantage whereas passenger mutations have negligible

effect on cellular fitness at a given time point during the process of cancer evolution and progression. Driver aberrations are generally required for tumour growth and contribute towards disease progression. The cancer driver aberrations can be identified based on most affected genomic regions which are frequently mutated, rearranged or reflecting copy number gains or losses in a particular cancer type (Wood *et al.*, 2007)(Lawrence *et al.*, 2014). The number of driver mutations required for cancer to reach malignancy is still debatable but is generally considered to be in between 2 and 20 in most solid cancers (Beerenwinkel *et al.*, 2007) whereas in the case of certain subtypes of Acute Myeloid Leukaemia (AML), it can be as low as one driver mutation (TCGA, 2013a). Difficulties in the identification of subclonal driver mutations due to spatial heterogeneity (Gerlinger *et al.*, 2012a) and their tendency to follow diverse genetic pathways makes it cumbersome to completely characterize the driver aberration landscapes (Wood *et al.*, 2007).

A large number of passenger mutations are associated with driver mutations. For example, studies conducted in AML have revealed that the majority of passenger mutations were present before the cell of origin was transformed into a cancerous cell (Welch *et al.*, 2012a). However, passenger aberrations can also be acquired during the process of cancer evolution and progression (clonal selection during evolution) (Gerlinger *et al.*, 2012a). Overall, the distinction between driver and passenger mutations during cancer may be dynamic as the fittest genotype may not be the same in all cancers or at all places or at all times because the selection is environment-dependent (Yap *et al.*, 2012). In certain cases, late-stage cancer may not always rely on an early driver event. As certain passenger events perhaps act as driver events in these cancer cells (Feldser *et al.*, 2010). However, for some driver mutations, environmental variation may matter less because such mutations almost invariably confer a selective advantage. For example, p53 mutation is found in various types of cancers however, they are not always the initiating mutations driving the cancer progression. They are sometimes selected once a particular stage of evolution has been reached (Rivlin *et al.*, 2011). In some cases, changes in environmental conditions may influence on which genetic changes are selected. This is evident from the fact that many cancer genes only drive the disease when they are mutated in the germline (Amberger *et al.*, 2015). These examples indicate that mutational landscape changes through space and time and the effect of individual mutation is context-specific.

1.2 Models of cancer evolution

To further understand the consequence of mutations, it is quite important to study the evolutionary trajectory of cancer cells at different time points during the process of disease progression. However, the central problem of such approaches is that patients cannot be ethically biopsied at multiple time points during the disease development. Analysing patient samples at a single time point leads to an incomplete picture of different stages of cancer progression specially if intermediate or transient clones are involved during the process. On the other hand, static analysis provides a snapshot of the disease development which can be taken at different time points during the course of evolution in order to develop a better understanding of the disease evolution. Another issue is associated with the fact that early stages of cancer are often silent clinically and either not detectable or detected by chance in apparently healthy individuals (Busque *et al.*, 2012; Laurie *et al.*, 2012; Genovese *et al.*, 2014; Martincorena and Campbell, 2015). Although the general representative models of cancer evolution are still debatable, following are the competing models which were broadly proposed (Davis, Gao and Navin, 2017) –

1. Linear Evolution (LE)
2. Branching Evolution (BE)
3. Neutral Evolution (NE)
4. Punctuated Evolution (PE)

LE is the most commonly proposed model of cancer evolution suggesting that mutations were acquired early in a stepwise manner consequently leading to more malignant stages of cancer (Fearon and Vogelstein, 1990). LE model posited that new driver mutations outcompete all previous clones by providing strong selective advantage (also known as selective or clonal sweep) during the process of cancer evolution (Fig 1.2A). Experimental evidence of LE was originally found in the case of X-inactivation in cancer where human cancer cells showed inactivation of single clonal X-allele throughout the cancer mass due to selection of dominant clone (Linder and Gartler, 1965). Another example of LE was further shown by Fearon and Vogelstein where their work suggested that colon cancer progresses through stepwise mutations in a linear order. This sequential acquisition of mutations further led to more

malignant stages of cancer growth (Fearon and Vogelstein, 1990). Overall, LE model suggests a stepwise acquisition of mutations consequentially leading to metastasis and more advanced stages of disease progression. However, most studies suggesting LE pattern of clonal selection lack the measurement of genome-wide markers, indicating that some of the heterogeneous mutations defining different clones during the process of disease progression may have been missed.

BE represents another model of cancer evolution where clones diverge from an ancestor and evolve in parallel during disease progression, leading to multiple clonal lineages (Fig 1.2B). In contrast to LE model of evolution, the selection sweep is not reported as the clones can have similar levels of fitness. BE has been reported in many cancer types, including, leukaemia (Gawad, Koh and Quake, 2014), breast cancer (Nik-Zainal, Van Loo, *et al.*, 2012), liver cancer (Ling *et al.*, 2015), colorectal cancer (T. M. Kim *et al.*, 2015), ovarian cancer (McPherson *et al.*, 2016), prostate cancer (Boutros *et al.*, 2015), melanoma (Harbst *et al.*, 2016) and brain cancer (Sottoriva *et al.*, 2013). Although these studies confirm the presence of BE model of cancer evolution, unfortunately, they differ in the number of clones reported. The number of such clonal populations in BE also varies among the patients having the same cancer. For example, a deep-sequencing based study conducted on over hundred triple-negative breast cancer patient samples identified 1 to 19 subclones per patient (Shah *et al.*, 2012). These kind of variability in the genomics data can be consequential to the number of cells sequenced, the location of the tumour from where cells were obtained or to the sequencing depth. A typical feature of BE includes clonal evolution, where a particular driver mutation gets accumulated gradually over the process of development and further leads to clonal expansion within the tumour. This kind of selection process during the continuous clonal evolution is supported by sub-clonal driver mutations or convergent evolution. Sub-clonal driver mutations have been usually reported from single cell analysis on multi-region sampling. For example, in case of a multi-region sequencing conducted on breast cancer patient cohort, 13 out of 50 patients reported the presence of sub-clonal driver mutations (Yates *et al.*, 2015). Another line of selection during clonal evolution can be seen in cases of convergent evolution (CE). CE is a model of cancer evolution where two independent lineages of cancer progression have mutation in the same driver gene, leading to different clonal expansions. Certain cases of lung cancer have provided evidence for CE where 5 out of 10 cases were found to follow CE during

the disease progression (De Bruin *et al.*, 2014). These examples clearly indicate that subclones can co-exist and expand in parallel to each other without outcompeting each other unlike in case of LE model of evolution.

NE represents another model of evolution which is basically an extreme case of BE. NE is based on the assumption that no selection or fitness change is observed during most of the lifetime of cancer evolution (Fig 1.2C). NE model of evolution basically hypothesizes that during the process of cancer progression, random mutations accumulate over the period of time leading to a genetic drift. NE was also proposed originally in the case of species evolution, strongly challenging the theory of natural selection laid by Darwin (Kimura, 1983). As discussed above in BE, natural selection is observed in cancer cells due to the presence of sub-clonal driver mutations and convergent evolution. However, there are certain cases of cancer evolution providing evidence for NE. For example, in a study involving The Cancer Genome Atlas (TCGA) data analysis conducted by Williams et al., a model of NE was applied to study the sub-clonal allele frequencies. The study provided an evidence that one-third of the examined cancer cases follow NE model of cancer evolution (Williams *et al.*, 2016). These studies indicate the presence of NE model of evolution; however, this evidence also have limitations associated with an identification of sub-clonal mutations perhaps due to poor sequencing quality or low sequencing depth or may also suggest that non-genetic mechanisms have a role during cancer evolution.

The models of evolution mentioned above are based on the assumption that mutations are acquired sequentially over the period of cancer progression. However, in certain cases of cancers, it has been observed that a large number of genomic aberrations may occur at very early stages of cancer progression, during short bursts of time period (Fig 1.2D). The term ‘Punctuated Evolution’ was proposed as ‘Punctuated Equilibrium’ by Gould and Eldredge in 1970 which challenged the longstanding paradigm of Darwinian evolution (Gould and Eldredge, 1993). PE evolutionary model is fundamentally different from other models of evolution as in this case, the assumption is based on the fact that cancer cells are pre-programmed at the earliest stages of cancer development and are destined to become invasive, metastatic or resistant to treatment. In contrast to the models mentioned above, which majorly

rely on point mutations, PE model of cancer evolution has been reported in studies indicating DNA copy number aberrations and chromosomal structural rearrangements.

PE has been evidenced in two groups of studies- one involving localized phenomenon on single chromosomes and the other leading to aneuploidy consequential to whole genome rearrangements. One such phenomenon involving localized chromosome rearrangements is known as 'chromothripsis'. Chromothripsis can be defined as a single catastrophic event involving shattering and reassembly of a chromosomal arm (Williams, Sottoriva and Graham, 2019) and is more specifically defined by many oscillating copy number states in which breakpoints map between adjacent segments on a single chromosome, and has been reported in many cancer types including colorectal cancer (Kloosterman *et al.*, 2011), prostate cancer (Berger *et al.*, 2011), etc. Another group of studies where PE has been found to be implicated is the genesis of genome-wide aneuploidy. In a study involving 57 patients of prostate cancer, a phenomenon called 'chromoplexy' was reported where genome wide translocations and copy number alterations were interdependent and occurred concurrently in short periods of time (Baca *et al.*, 2013). Overall, these studies indicate that chromosomal rearrangements and copy number alterations may evolve through PE model of tumour progression.

Ultimately, all these models are variations of the same, nuanced by the initial, detectable event where either the clone gain a selective advantage or diverge from the ancestral clone during the process of evolution.

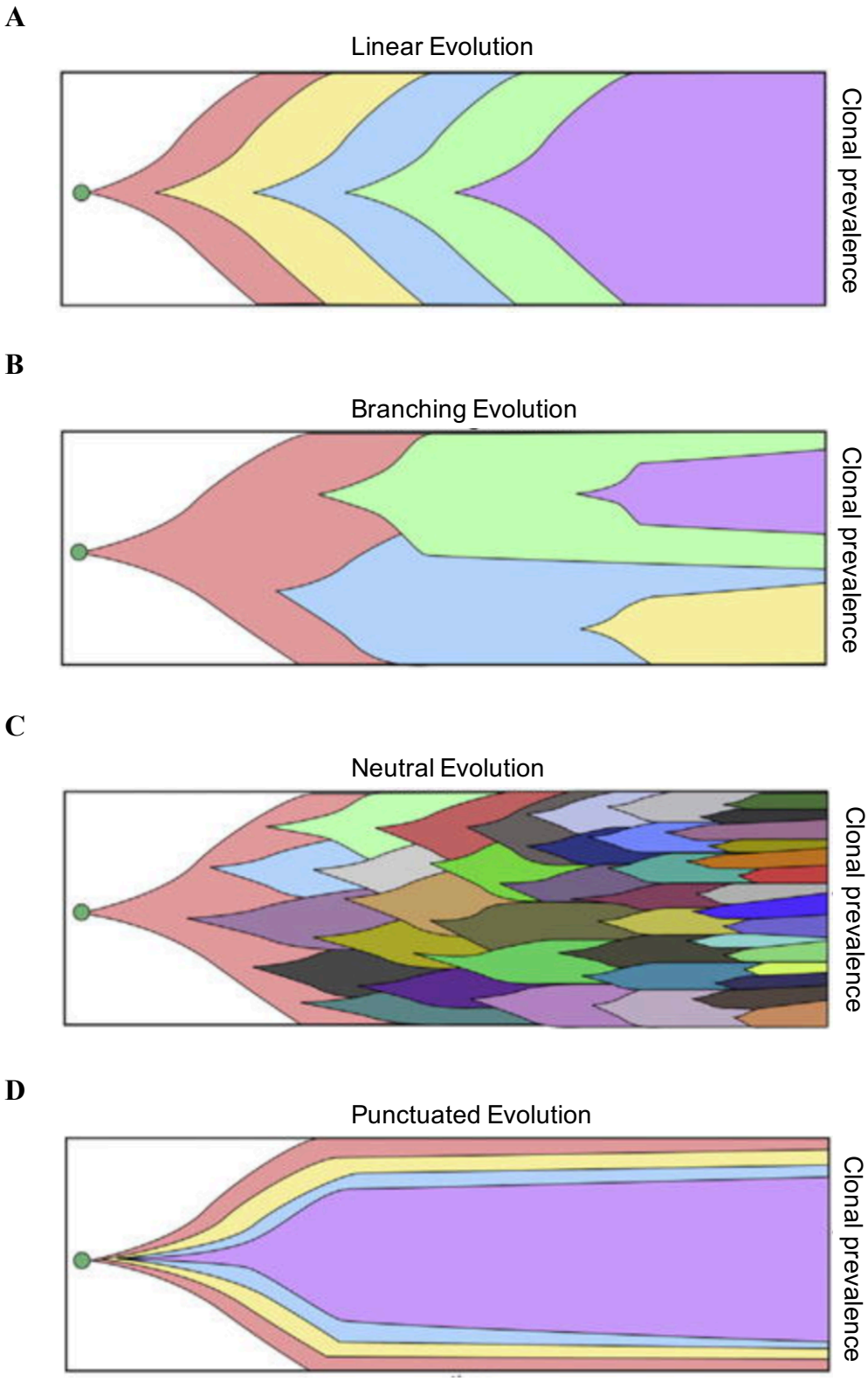

A
Linear Evolution
Clonal prevalence
B
Branching Evolution
Clonal prevalence
C
Neutral Evolution
Clonal prevalence
D
Punctuated Evolution
Clonal prevalence

Figure 1.2: Models of tumour evolution.
Illustration of tumour evolution models showing dynamic changes in clonal frequencies over time. This figure is based on the original study (Marusyk and Polyak, 2010) and derived from (Davis, Gao and Navin, 2017). **(A)** Linear Evolution **(B)** Branching Evolution **(C)** Neutral Evolution **(D)** Punctuated Evolution. Colours indicate clones with different genotypes.

1.3 Mechanisms driving cancer evolution

1.3.1 Genetic variability

One of the mechanisms underlying clonal evolution of cancer is the continuous acquisition of somatic mutations contributing towards genetic variability. This genetic variability could be allelic or locus-specific. Allelic variability represents different variants at a single gene locus leading to similar phenotypic expressions of disease whereas locus-specific variability occurs when variants at different gene loci lead to similar disease phenotype. Genetic variability may be consequential to development of a genetically distinct subclonal population of cells followed by selective outgrowth of clones having a phenotype advantage within a given microenvironment as discussed above (Nowell, 1976). In case of selective sweep, as mentioned above, whereby a new clone takes over the entire population by replacing any ancestral clones, leads to a homogeneous cell population. In case of LE, if a new clone fails to outcompete its predecessor, a degree of genetic variability will be observed (Welch *et al.*, 2012b). However, in case of BE, the evolution of subclonal population in parallel results in extensive subclonal diversity. In certain cases of Acute Lymphoblastic Leukaemia (ALL) following BE, a heterogeneity was observed between leukaemia propagating cells (Anderson *et al.*, 2011). However, the phenomenon of genetic variability has a clearer topography in solid tumours. In certain cases of clear-cell renal cell carcinoma which followed branched evolutionary trajectory, spatial separation of tumour subclones was observed. These spatially separated subclones were found to harbour heterogeneous somatic mutations and copy number events leading to genetic variability (Gerlinger *et al.*, 2012b). In certain cases of breast cancer (Shah *et al.*, 2009), pancreatic cancer (Campbell *et al.*, 2010) and medulloblastoma (Wu *et al.*, 2012) there is evidence for clonal diversity between primary and metastatic site. The genetic variation between primary and metastatic tumour sites may be attributed to the selection pressure imposed by distinct micro-environmental niche such that each subclone occupying their

individual niche evolve independently relative to each other (Junttila and De Sauvage, 2013). These intermingled heterogeneous clones obtained from the same biopsy may display amplification of different proteins, some of which regulate key signalling processes and whose dysregulation may impact cancer development, for example, receptor tyrosine kinases (Snuderl *et al.*, 2011). The genetic diversity observed at this intercellular level may be a consequence of genomic instability. This genomic instability leads to intercellular variability which increases phenotypic variation by broadening the pool of cells subjected to selection and further leading to the emergence of complex subclonal architecture (Cahill *et al.*, 1999).

The Deoxyribonucleic Acid (DNA) replication machinery in a genome works at a high precision where the genome is replicated and divided with high fidelity. This high accuracy rate results in a very low endogenous mutation rate in somatic cells which is estimated to be $0.77*10^{-9}$ per site per cell division (Lynch, 2010). Similar to this observation, the errors in chromosome aggregation are also very rare which happen roughly at a rate of 1 per 100 cell divisions and poorly tolerated in non-transforming cells (Thompson and Compton, 2008). Disruption of mechanisms which maintain the integrity of the genome or exposure to any form of exogenous mutagens will elevate the genomic aberration rate. Most of the solid tumours as well as haematopoietic malignancies are found to have at least one form of genomic instability (Gordon, Resio and Pellman, 2012). Despite the clear association of genomic instability and cancer, the fact whether elevated mutation rates are required for cancer evolution, still remains debatable (Tomlinson, Novelli and Bodmer, 1996). The mutational burden within a large population of cancer cells is likely to be extensive even in the presence of normal somatic mutation rate (Tomlinson, Novelli and Bodmer, 1996). In accordance with this, certain studies supported by mouse models have revealed that genomic instability increases the risk of cancer progression (Weaver *et al.*, 2007).

Different genetic lesions ranging from point mutation frequency to small insertions and deletions, large-scale chromosomal rearrangements and alterations in ploidy may lead to genomic instability (Table 1.1). The different genetic lesions may have distinct phenotypic penetrance, for example, most of the point mutations are neutral whereas a chromosome gain or loss may likely have functional consequences. As different patterns of genomic instability lead to distinct genomic footprints, it is possible to interrogate how genomic instability has

shaped tumour evolution using sequencing technology (discussed later) and copy-number data analysis (Nik-Zainal, Alexandrov, *et al.*, 2012). A recent study provided a deep insight into 30 different mutations in different cancer types and unravelled 20 distinct signatures of process that mutate DNA (Alexandrov *et al.*, 2013). In addition to the observation of mutational patterns impacting global genomic architecture, specific mutations may occur at higher frequency in the context of a particular instability mechanism. One such example is lung cancer where smoking increases mutation rate and the frequency of C.G → A.T transversions. This observation led to the inference that mutations in *TP53* can be attributed to DNA damage from cigarette smoke carcinogens (Pfeifer *et al.*, 2002).

The increase in the frequency of point mutations may be attributed to different exogenous factors (Alexandrov *et al.*, 2013) along with defects in DNA repair pathways including, base-excision repair and mismatch repair (Shah, Hile and Eckert, 2010). Apart from seeing this hypermutation effect, mismatch repair is also associated with expansion and contraction of repetitive tracts of DNA (microsatellite instability), leading to an increased frequency of frameshift mutations (Shah, Hile and Eckert, 2010). Different mechanisms impacting gross chromosome structure, such as DNA replication stress (Nikolaev *et al.*, 2012), defects in homologous recombination and domains that loop together in the interphase nucleus may also lead to an increase in point mutation frequency (Yang *et al.*, 2011).

Different mechanisms of genomic instability can contribute towards distinct pattern of point mutations- replication stress is associated with an enrichment of mutations in large genes (Nikolaev *et al.*, 2012) whereas microsatellite instability, depending upon the nucleotide composition of a particular set of genes, may affect specific genes more frequently than others (Shah, Hile and Eckert, 2010). A recent study has highlighted the importance of considering variation in mutation rate across the genome when identifying putative driver mutations where the authors have found that the mutation frequency across the genome is strongly correlated with DNA replicating timing and transcriptional activity (Lawrence *et al.*, 2013a).

One of the most common forms of genomic instability is chromosomal instability which occurs at high frequencies across many cancer types and is associated with aggressiveness, drug resistance and poor prognosis (McGranahan *et al.*, 2012). Chromosomal instability refers to

any type of karyotypic abnormalities which may not be restricted to translocation, inversions and deletions and may also include an increased rate of change in chromosome number or structure. A certain proportion of chromosomally unstable tumours also show increased ploidy, suggesting that genome doubling could act as precursor to chromosomal instability (Ganem, Storchova and Pellman, 2007). From a mechanistic point of view, chromosomal instability has been found to be driven by mitotic defects in chromosome attachment which further leads to DNA damage and structural rearrangements by entrapment of chromosomes during cytokinesis (Janssen *et al.*, 2011) or through aberrant replication or fragmentation of micronuclei (Crasta *et al.*, 2012). On the other hand, structural chromosomal aberrations can also compromise the accuracy of chromosome aggregation during mitosis (Chan, North and Hickson, 2007). Overall, the impact of structural chromosomal aberrations generated from mitotic and pre-mitotic defects, on chromosomal instability in cancer still needs to be explored.

It is worth noting that a high number of mutations or chromosomal rearrangements do not necessarily imply ongoing genomic instability but could reflect a historical prolonged period of mutational acquisition or the occurrence of transient mutational bursts (Stephens *et al.*, 2011). The evolution of cancer does not only proceed gradually through clonal selection of point mutations or chromosomal alterations but could also be subjected to certain punctuated phases like chromothripsis as described above. Another such example is telomere dysfunction, catalysed by eroded chromosome ends, contributing towards genomic instability (Artandi and DePinho, 2009). These punctuated events are likely to contribute towards cancer evolution. The observations of active DNA damage response and variation in chromosome copy number, suggest that ongoing instability occurs in a substantial proportion of tumours (Gorgoulis *et al.*, 2005).

There are increasing number of studies suggesting the requirement of identification of low-frequency genetically and functionally distinct subclones at diagnosis (Landau *et al.*, 2013). One such study conducted on Chronic Lymphocytic Leukaemia (CLL) suggested the presence of distinct subclones that harbour distinct driver events during therapy. In this study, the patients harbouring a particular subclonal driver in their pre-treatment samples showed a shorter time to re-treatment or death as compared to the patients without any evidence of subclonal driver events (Landau *et al.*, 2013). Such studies highlight the importance of

longitudinal tumour sampling over the disease course and throughout the treatment as the subclonal driver event may not be detectable in a single biopsy or even in multiple biopsies collected before treatment. The ultimate challenge is in identifying the subclones with prognostic power that bias therapeutic choices.

It is also worth appreciating that phenotypic divergence cannot be solely attributed to genetic variability. In certain cases, genetically homogeneous subclones have been found to have distinct functional trajectories after exposure to chemotherapy, sometimes dependent on the presence of quiescent cells surviving the cytotoxic pressure (Kreso *et al.*, 2013a). Distinct phenotypic outcomes cannot be solely determined through genetic variability between subclones, but also through stochastic events in gene expression and protein stability as well as epigenetic divergence and microenvironmental fluctuations (Kreso *et al.*, 2013a).

1.3.2 Epigenetic variability

While genetic variability has been the most studied mechanism of clonal evolutionary development of cancer, it is widely recognized that epigenetic modifications are likely to play an important role in ultimately affecting fitness (Clark and Melki, 2002) (Baylin and Jones, 2011). The presence of phenotypic variation in genetically uniform cells, in terms of essential properties such as survival capacity and proliferative potential, strengthens the existing epigenetic variation (Kreso *et al.*, 2013b) (Spencer *et al.*, 2009). The term epigenetics originally referred to the stable heritability of a phenotype that results from changes in chromosome without any alterations in DNA sequence (Berger *et al.*, 2009). However, a more widespread usage of the term also includes DNA methylation, histone modifications, the transcriptional effects of RNA interference and nuclear organisation (Bird, 2007). Epigenetic mechanisms achieve diverse stable phenotype in genetically identical cells by controlling the abnormal transcriptional activation/repression of genes, referred to as ‘Epimutation’ (Holliday, 1987). These epimutations are heritable and together with genetic landscape, affect the process of cancer evolution (Richards, 2006).

Based on the range of autonomy of epigenetic variability in relationship to genotype context, it can be divided into three different categories (Fig 1.3)-

1. Obligatory epigenetic variation
2. Facilitated epigenetic variation
3. Pure epigenetic variation

Obligatory epigenetic variation (Fig 1.4A), as the name indicates, is completely dependent on genetic variation. In this case, there is a strict one-to-one correspondence between the epigenotype and either cis- or trans-acting genetic variation (Richards, 2006). Some of the examples include transposon-associated alleles where the presence of transposon is strictly associated with epigenetic silencing of coding sequence present in its vicinity (Lippman *et al.*, 2004).

Facilitated epigenetic variation (Fig 1.4B), on the other hand, is more of genotype directed mechanism where genotype potentiates the epigenotype in a probabilistic but not deterministic manner. Examples include the formation of epialleles in DNA methylation mutant background in case of *Arabidopsis thaliana* (Kakutani *et al.*, 1996).

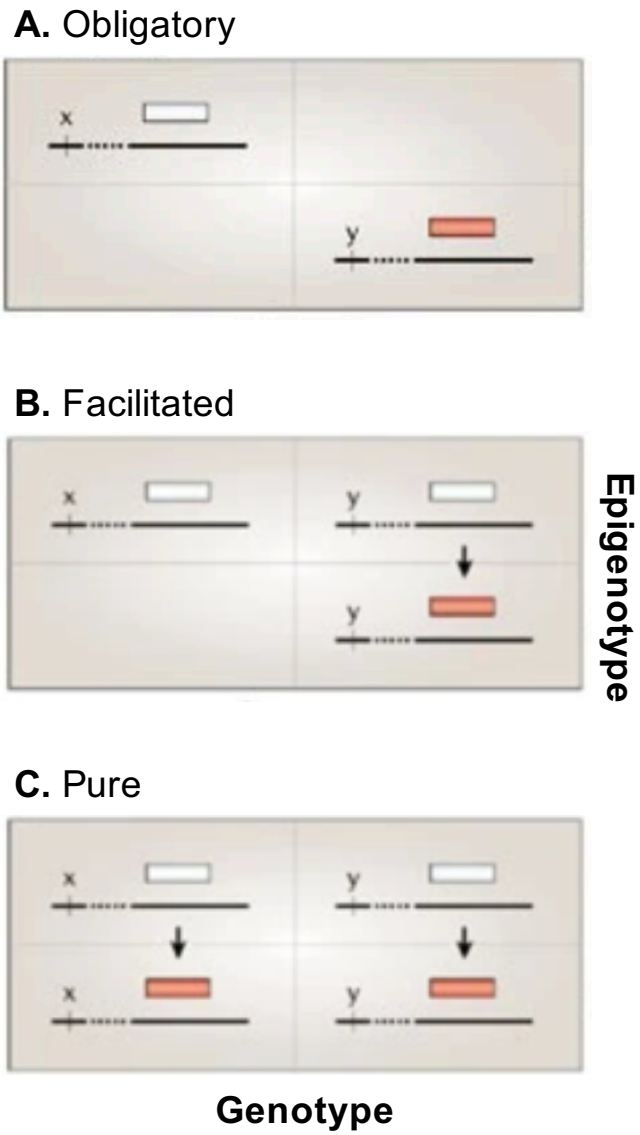


Figure 1.3: Types of epigenetic variability bases on relationship with genotype.
The horizontal axis distinguishes between two genotypes that are represented by alleles x and y. At a genomic location either in cis or in trans (dashed line), two alternative epigenetic states are depicted as either open or filled red boxes. **(A)** Obligatory type, where the epigenotype of the locus is strictly determined by genotype, **(B)** Facilitated epigenetic variation (epiallele formation: open box–red box) that can occur, in a probabilistic manner, only in the context of genotype y, **(C)** Pure phenotype, where stochastic events generate alternative epialleles at some finite frequency regardless of the genotype. Adapted from (Richards, 2006).

Pure epigenetic variation (Fig 1.4C), however, is generated by stochastic events which are largely independent of genetic variations. These stochastic events may include random errors

in propagating silent genetic states after DNA replication and the targeting of silent chromatin to ectopic genomic locations. One such striking example with consistent stochastic alterations leading to epigenetic variation is the growing divergence in epigenotype in monozygotic twins during ageing (Fraga *et al.*, 2005).

The very first observation of the presence of an epimutation in genotype context leading to cancer was aberrant methylation of the maternal allele of the imprinting control region (ICR) of the *H19* gene in the normal kidney parenchyma of patients with Wilms tumour. This resulted in the loss of imprinting of the closely linked, reciprocally imprinted, H19 and insulin-like growth factor 2 (IGF2) genes (Moulton *et al.*, 1996). Another example of epimutation was *MLH1* in Lynch syndrome, which was the first epimutation to be observed in a non-imprinted gene. In this case, an aberrant methylation in *MLH1* promoter was observed, accounting for 1–10% of cancer patients who meet the clinical diagnostic criteria for Lynch syndrome with a loss of *MLH1* expression in their tumours. *MLH1* epimutations were characterized by soma-wide monoallelic methylation of the *MLH1* CpG island promoter accompanied by allelic transcriptional inactivation (Hitchins *et al.*, 2007).

These examples indicate the presence of locally disordered methylation patterns which is one of the epigenetic mechanisms leading to alternative epigenetic states known as epialleles, thereby, contributing towards epigenetic variability. Although global hypomethylation of cancer was described in early 1980s, with frequent focal hypermethylation of key regulatory regions (Jones, 1999), recent literature has indicated the presence of stochasticity in DNA methylation patterns (Landau *et al.*, 2014). Studies conducted in diffuse large B-cell lymphoma (DLBCL) associate the presence of differential methylation patterns with relapse and indicate that the epigenomic variation in the form of heterogeneous methylation pattern can be used as a predictive marker in case of DLBCL (Pan *et al.*, 2015).

Another study performed on chronic lymphocytic leukaemia (CLL) suggested the presence of stochastic patterns of DNA methylation during the process of disease evolution. The increase in locally disordered methylation was found to be associated with adverse clinical outcome. The authors also proposed that stochastic methylation alterations enhance epigenetic plasticity and further enable tumour cells to better explore the evolutionary space in search of superior

fitness trajectories (Landau *et al.*, 2014). They proposed a model associating epigenetic landscape with genetic diversification where cellular populations with a preserved epigenetic landscape showed a limited capability of genetic diversification. However, the population of cells with a more malleable epigenetic landscape gave birth to new subclones leading to enhanced genetic diversity and poor clinical outcome.

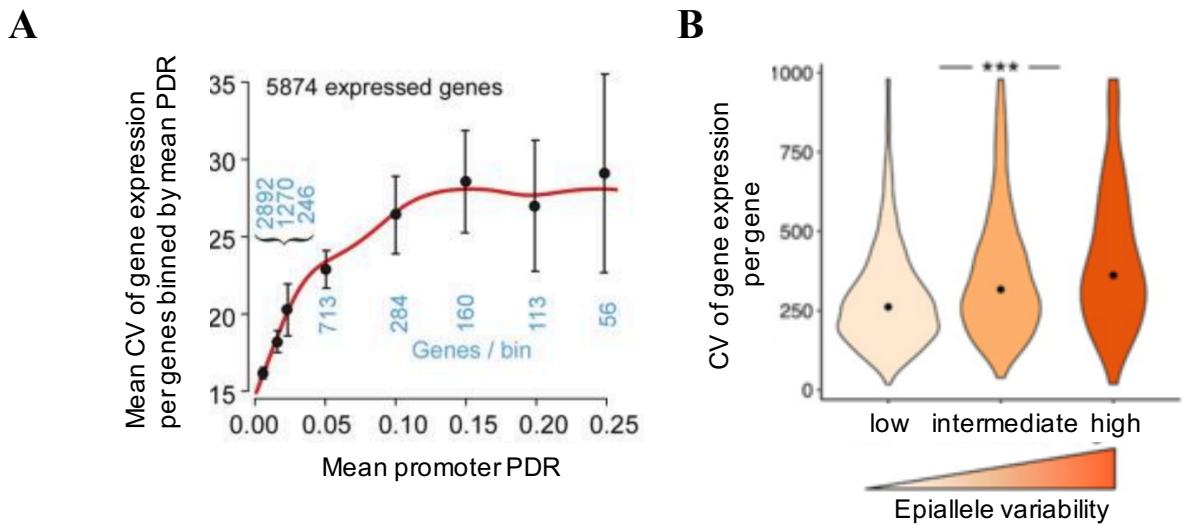


Figure 1.4: Relationship between epigenetic variability and transcriptional variability.
(**A)** PDR and expression variability as measured with coefficient of variation (CV) of 5,874 transcribed genes. Black circles (brackets)-- mean CV (95% CI) for genes within PDR bins (number of genes per bin in blue). Red line - cubic smoothing spline of CV and PDR values (unbinned) (modified from (Landau *et al.*, 2014) , (**B)** Violin plots of transcript expression level variance as measured by single cell RNA-sequencing (AML_130 relapse sample) and association (ANOVA test, $p < 2.2\times1016$) with low (< 0.05), intermediate (0.050.2) and high (0.21) epiallele shift within respective gene promoters. Wilcoxon signed rank tests and ANOVA test: $p < 0.001$*** (modified from (Li *et al.*, 2016)).

It is worth noting that locally disordered methylation is associated with variability in transcriptional landscape with a decoupling of the relationship between promoter methylation and gene expression (Landau *et al.*, 2014) (Fig 1.4A). This observation was compatible with another study conducted on AML where the genes downstream of the promoters associated with epiallele shifts, also showed variability in gene expression at single-cell level (Li *et al.*, 2016) (Fig 1.4B). These observations imply that apart from genetic and epigenetic variability playing a substantial role in determining the phenotypic variability, the consequential

transcriptional variability may also contribute mechanistically during the process of cancer evolution.

1.3.3 Transcriptional variability

The concept of cell-to-cell transcriptional variability was first highlighted by Novick and Weiner in 1957, who showed that the production of beta-galactosidase in individual cells was highly variable and random, where the process of induction led to an increase in the proportion of cells expressing the enzyme rather than increasing every cell's expression level equally (Novick and Weiner, 1957). The cell-to-cell transcriptional variability reflects the variability in the number of mRNA molecules produced from a given locus at a certain time point (Ko, 1991) (Mcadams and Arkin, 1997). This is important for stochastic fate specification which complements lineage- and signalling-based mechanisms to further diversify cell types during development (RJ and C, 2010). However, to date, the mechanisms contributing towards transcriptional variability still remain under exploration (Urban and Johnston, 2018).

The very first attempt at understanding the mechanisms behind transcriptional variability was made by Elowitz et al. who introduced the concepts of intrinsic and extrinsic sources of transcriptional variability (Elowitz *et al.*, 2002). For this, the authors quantified the variability in the expression from a promoter in *E. coli* by introducing two copies of the same promoter into the genome of *E. coli*, one driving the expression of cyan fluorescent protein (CFP) and the other driving the expression of yellow fluorescent protein (YFP). The authors suggested that extrinsic fluctuations are those that affect the expression of both copies of the gene equally in a given cell (Fig 1.5A). However, intrinsic fluctuations are those due to the randomness inherent to transcription and translation and thus should affect each copy of the gene independently (Fig 1.5A), adding uncorrelated variations in levels of CFP and YFP (Fig 1.5B). The authors found that both sources of noise can be significant depending on the promoter, the key regulatory region in bacteria. The observation that the time scale for intrinsic noise fluctuations is much shorter (~9 minutes) than that for extrinsic noise (~40 minutes) suggests that extrinsic noise may affect cellular phenotypes more strongly than intrinsic noise, at least in *E. coli* (Rosenfeld *et al.*, 2005). These findings were further confirmed by Ozbudak et al.

who also quantified noise in gene expression in the prokaryote *Bacillus subtilis* (Ozbudak *et al.*, 2002).

These studies highlight that gene expression variability can be majorly subdivided into two types-

1. Extrinsic variability
2. Intrinsic variability

Extrinsic gene expression variability arises from heterogeneous environmental conditions surrounding the cell (Chalancon *et al.*, 2012). It may also be due to asymmetric cell divisions or variability in abundance of gene expression machineries like cell-to-cell variation in ribosome abundance or any changes in the local distribution and concentration of general transcription factors or other proteins. Extrinsic variability varies from cell-to-cell but has the potential to affect all genes and is therefore a gene-independent characteristic.

Conversely, intrinsic source of gene expression variability can be attributed to variation in chromatin modifier binding at individual gene loci and binding of transcription factors (Chalancon *et al.*, 2012). Other factors which may impact intrinsic variability include promoter architecture, rate of protein synthesis/ degradation inside a cell as well as transcriptional bursting (discussed later). Thus, intrinsic variability is considered a gene-dependent characteristic because its effects vary from gene-to-gene and cell-to-cell.

One of the major reasons for gene expression variability is the occurrence of transcription in the form of transcriptional ‘burst’ (Cai, Friedman and Xie, 2006). Transcriptional burst is the phenomenon by which mRNA is produced from the promoter of a gene, in the form of bursts, switching between periods of prolonged inactive (or ‘off’) states and short-lived active (or ‘on’) states (Thattai and Van Oudenaarden, 2001). The perturbation in dynamics and frequency of the pre-initiation complex (PIC) assembly is one of the reasons for transcriptional bursting (Raj and Van Oudenaarden, 2009). Transcriptional bursting is one of the major mechanisms contributing towards intrinsic variability and majorly depends on the following-

1. TATA-box- a DNA binding motif present in the gene-promoters. The variation in the actual DNA sequence of the motif impacts the stability of PIC and hence contributes towards transcriptional bursting (Lam, Steger and O'Shea, 2008; Corrigan *et al.*, 2016; Ochiai *et al.*, 2020).

2. Nucleosome occupancy and positioning- in a given population of cells, the nucleosome occupancy and positioning at a particular genomic coordinate influences transcriptional bursting (Blake *et al.*, 2003).

3. Transcriptional pausing- the presence of polymerase pause sites may impact bursting by either stalling polymerases or causing premature termination (Ribeiro *et al.*, 2010)(Larsson *et al.*, 2019).

4. Chromatin epigenetics- Histone modifications and DNA methylation can be added and removed in a switch-like manner and may impact bursting kinetics (Lim and Van Oudenaarden, 2007; Miller-Jensen *et al.*, 2011; Larsson *et al.*, 2019)

One of the models which is majorly studied to understand the mechanism of transcriptional bursting is called 'two-state' model (Peccoud and Ycart, 1995). The model is based on the observation that transcriptional bursting occurs when a gene promoter fluctuates between an "on" and "off" state for different periods of time. Each time the promoter switches to an "on" state, a burst of transcription is produced (Kumar, Singh and Kulkarni, 2015). The majority of intrinsic gene expression variability arises from stochastic production of mRNA due to the random binding of transcription factors and other transcriptional machinery to the DNA, along with the turnover of mRNA and protein at a certain point of time (Raser and O'Shea, 2004a). The rate of binding at the promoter or enhancer is termed as K_{on}, whereas rate of dissociation is termed as K_{off} (Fig 1.6A). These rates of binding and dissociation are majorly dependent on protein availability and the presence of other bound proteins. Extrinsic fluctuations produce variability in the local concentrations of proteins, leading to alterations in the probability of protein binding over time. Gene expression variability also arises from variation in the transcription initiation rate (μ) and mRNA degradation rate (δ). The frequency, duration, and

amplitude of bursts determine the total amount of RNA produced from a particular gene (Fig 1.6B).

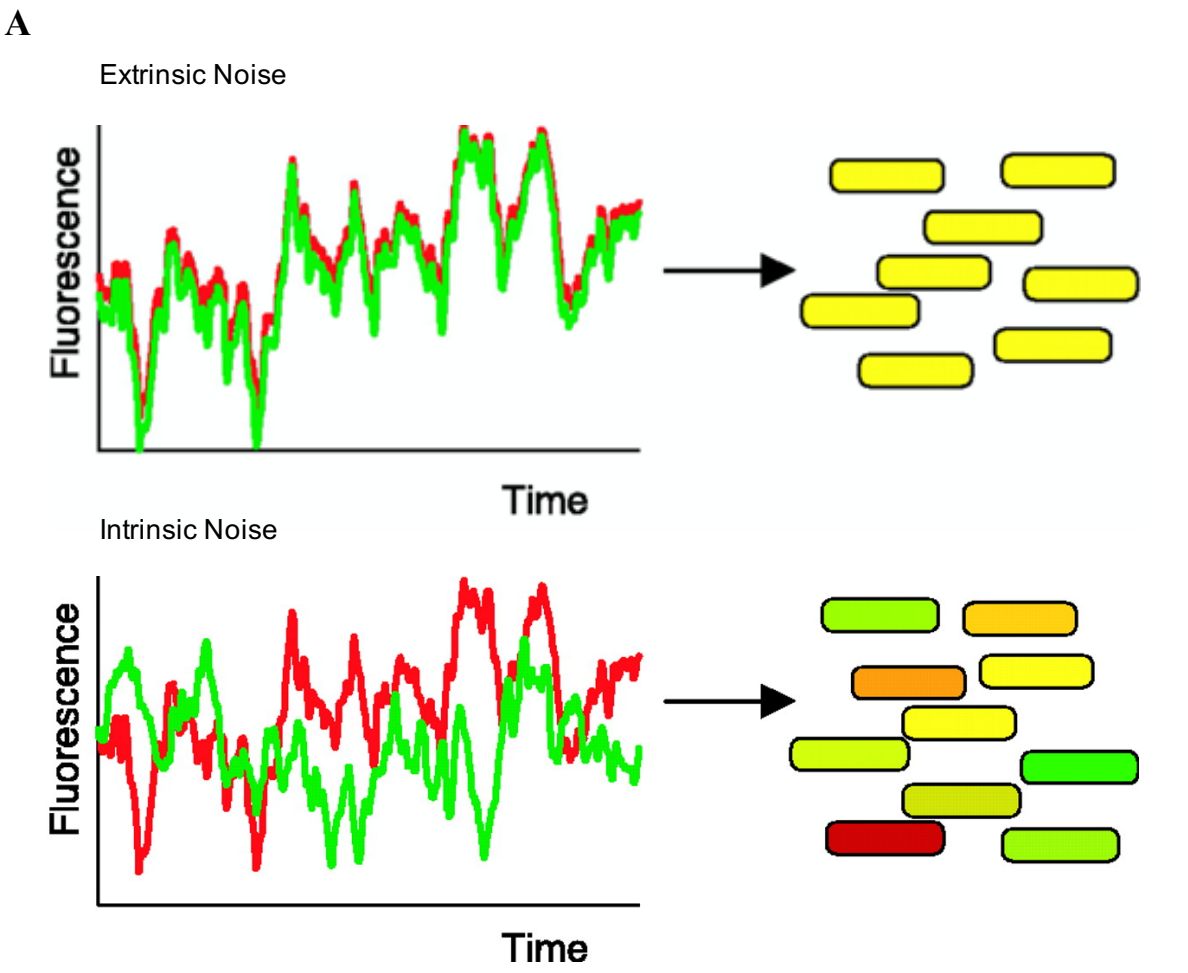


Figure 1.5: Intrinsic and extrinsic contributions to gene expression variability.
(A) Schematic depiction of the temporal behaviours of extrinsic noise (upper) and intrinsic noise (lower). Adapted from (Elowitz *et al.*, 2002), **(B)** Expected cell-to-cell variations when fluctuations are intrinsic, extrinsic or both. Adapted from (Raj and van Oudenaarden, 2008).

While the two-state model describes the phenomenon of transcriptional bursting and the gene-independent mechanisms contributing towards this, it does not take the gene-architecture into consideration. As mentioned above, The TATA box is a critical determinant of gene expression variability, where genes lacking a TATA box are associated with lower variability, while the presence of a TATA box promotes higher variability (Blake *et al.*, 2006a). In addition to promoter architecture, the interactions between enhancer-promoter also act as gene-dependent regulators of transcriptional bursting. One of the mechanisms by which enhancers may affect bursting is through enhancer-promoter looping (Fukaya, Lim and Levine, 2016a). The fluctuations in the kinetics of the cellular environment, such as changes in the concentration and distribution of transcription factors, change the probability of enhancer-promoter contact from cell-to-cell. These alterations in enhancer-promoter interactions lead to differences in bursting between cells (Tantale *et al.*, 2016).

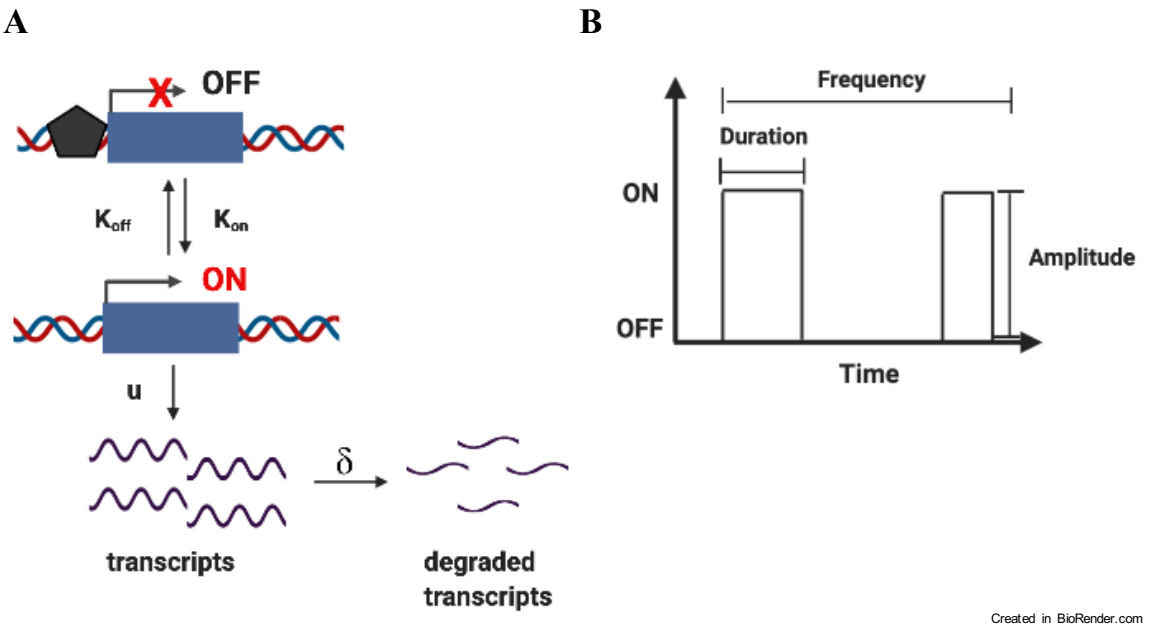


Figure 1.6: Two-state model of transcriptional bursting.
(A) The two-state bursting model suggests that a gene switches between an 'inactive' and 'active' state at a particular rate, K_{on} and K_{off}. In the 'active' state RNA transcripts are produced at a rate (u) and degraded at a rate (δ), **(B)** Transcriptional bursting parameters where each burst occurs for a particular duration (period of time) with a distinct amplitude (strength) and at a particular frequency. Modified from (Urban and Johnston, 2018).

Apart from promoter architecture and enhancer-promoter looping, the chromosome organization in topologically associated domains (TADs) may also be associated with the phenomenon of transcriptional bursting. Since these TADs facilitate DNA binding with proteins like cohesins and CTCF, the depletion of these proteins may subsequently lead to dysregulation of gene expression (Dixon, Gorkin and Ren, 2016). Literature suggests that intra-TAD CTCF binding promotes and stabilizes enhancer-promoter interactions which further result in a decrease in transcriptional bursting and variability (Ren *et al.*, 2017a). In addition to this, certain histone modifications, including H3K27ac as well as H3K36me3 have been shown to be positively correlated with burst frequency (Larsson *et al.*, 2019; Ochiai *et al.*, 2020). These examples shed some light on how gene-specific regulatory mechanisms in combination with transcriptional bursting may play a significant role in contributing towards transcriptional variability. It's worth understanding how this variability in gene expression may be consequential to stochastic cell fate specification further leading to diversification of cell fates.

One such example highlighting stochastic gene expression linked to cell fate decision is photoreceptor expression in Drosophila eyes. The Drosophila eye consists of a large number of optical units called ommatidia, each of which contains two cells that in turn express one of the specific pairs of photoreceptors, either Rh3 and Rh5 (for blue sensitive ommatidia) or Rh4 and Rh6 (for yellow sensitive ommatidia). Wernet *et al*., showed that this decision is almost exclusively due to the stochastic expression of the spineless gene during mid-pupation, with stochastically large levels of spineless expression resulting in the adoption of the yellow fate in roughly 70% of the ommatidia (Wernet *et al.*, 2006).

Another example of stochastic cellular differentiation is haematopoietic system where progenitor stem cells differentiate into different types of blood cells (Enver, Heyworth and Dexter, 1998). Chang et al., further delved into the link between stochastic differentiation and

gene expression variability. The authors showed that the fates of haematopoietic cells are influenced by the expression levels of the cell surface receptor Sca-1. Noise in the levels of Sca-1 partitions these cells into high-expressing erythroid fates and low-expressing myeloid fates (Chang *et al.*, 2008). These few examples further strengthen the notion that stochastic cell fate choices can be made by a particular cellular system in the presence of the source of transcriptional variability. With advances in genomics and single-cell studies, it is now possible to understand the role of transcriptional dynamics underlying cell-to-cell variability in different biological mechanisms including development and disease progression.

1.4 Single-cell technology as a means to quantify cell-to-cell transcriptional variability

As discussed in the previous section, transcription of a gene is carried out inside a cell in the form of transcriptional bursts, (Raj and van Oudenaarden, 2008), and individual genes may exhibit different pattern of burst kinetics. This difference in the pattern of burst kinetics contributes towards variability in gene expression, between individual cells of a population, both in terms of which genes they express and the level of expression. In order to understand the stochasticity or mechanistic regulation of the cell-to-cell transcriptional variability, it is quite important to understand the coregulation of transcription factors (TFs) and their expression level in a given population of single cells. Population studies did not only contribute towards masking of molecular and functional variability but also enforced an assumption towards individual cell behaviour by extrapolating the bulk studies data to single-cell level. This extrapolation of population studies suggested that all cells of a population express similar levels of a given gene (Fig 1.7A) and that co-expression of multiple genes in a given population of cells corresponds to co-expression in the individual cells (Fig 1.7B). However, studies conducted to date have indicated that often neither of these assumptions hold true at the single-cell level (Bengtsson *et al.*, 2005).

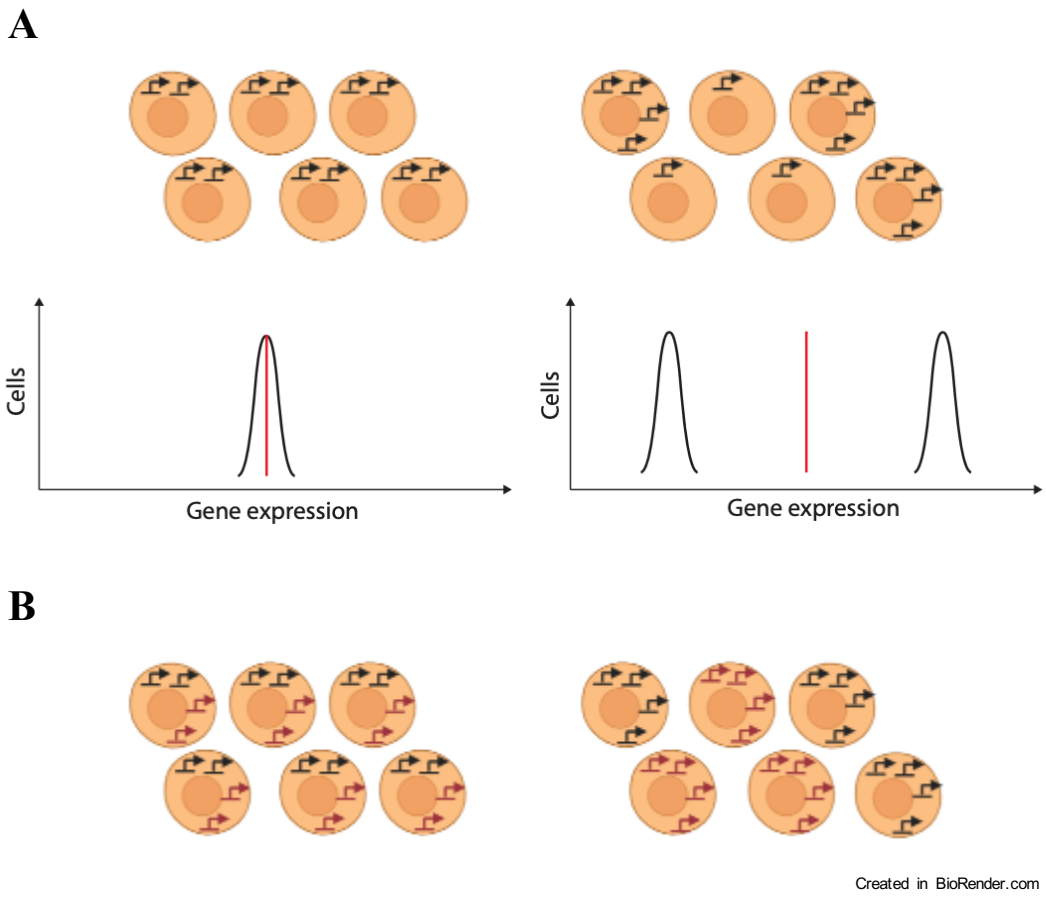


Figure 1.7: Single cell analysis reveals variability in gene expression patterns.
(A) Single cell analysis can distinguish whether all cells of a population express a similar level of a transcript (top left) or whether a small number of cells account for most of the expression (top right), which cannot be determined from population studies. In single cell studies, a homogeneous population would give a single expression distribution (bottom left) while a heterogeneous population would give a broader distribution, or multiple distributions (bottom right). In population studies, both sets of cells would seem to have the same level of expression (red lines), **(B)** Single-cell analysis can reveal whether co-expression observed at the population level actually occurs within the same single cells (left) or not (right). (Adapted from (Moignard and Göttgens, 2014)).

The first evidence of sequencing transcriptome at single-cell level was mentioned by Eberwine et al., (Eberwine *et al.*, 1992) along with Iscove and colleagues (Brady, Barbara and Iscove, 1990) where they pioneered the expansion of complementary DNAs (cDNA) from a single-cell using linear amplification by *in vitro* transcription and exponential amplification by PCR, respectively. These methodologies were further applied at a commercial scale level by making use of high-density DNA microarray chips and were subsequently adapted for single-cell RNA

sequencing (scRNA-seq) (Klein *et al.*, 2002)(Kurimoto *et al.*, 2006). Guided by these studies, the first description of scRNA-seq based on a next-generation sequencing platform (NGS) was available in the year 2009, where the study described the characterization of cells from early developmental stages (Tang *et al.*, 2009) (Fig 1.8). Since then several studies have proven the potential of scRNA-seq technology to identify the rare subset of population of cells which may have potential implications for furthering our understanding of drug resistance and relapse in cancer treatment (Shaffer *et al.*, 2017).

Since the first description of scRNA-seq by Tang and colleagues, a massive expansion in method development has been witnessed (Table 1.1). Such efforts are crucial as each method has distinct advantages and applicability. ScRNA-seq, in general, is composed of four steps (Hedlund and Deng, 2018)-

1. Single-cell isolation, capture and lysis
2. Reverse transcription
3. cDNA amplification
4. Sequencing library preparation

Isolation of single cell is the first critical step for obtaining the transcriptome information from an individual cell. One of the most commonly used technique is, fluorescence activated cell sorting (FACS), which has become quite a popular technique to isolate single-cells (Julius, Masuda and Herzenberg, 1972; Hwang, Lee and Bang, 2018). This method utilizes a fluorescent tagged antibody with which cells are tagged. These antibodies further recognize specific surface markers and enable sorting of distinct populations. In this case, cells are isolated magnetically, based on predetermined fluorescent parameters where a charge is applied to a cell of interest using an electrostatic deflection system. The same technique can be used for index sorting, widely used for single-cell based studies. However, FACS requires large starting volumes which makes it difficult to isolate cells from low input numbers (<10,000). Another method for isolation of single-cell from a solid sample is laser capture microdissection, which utilizes a laser system aided by a computer system to extract single-cells.

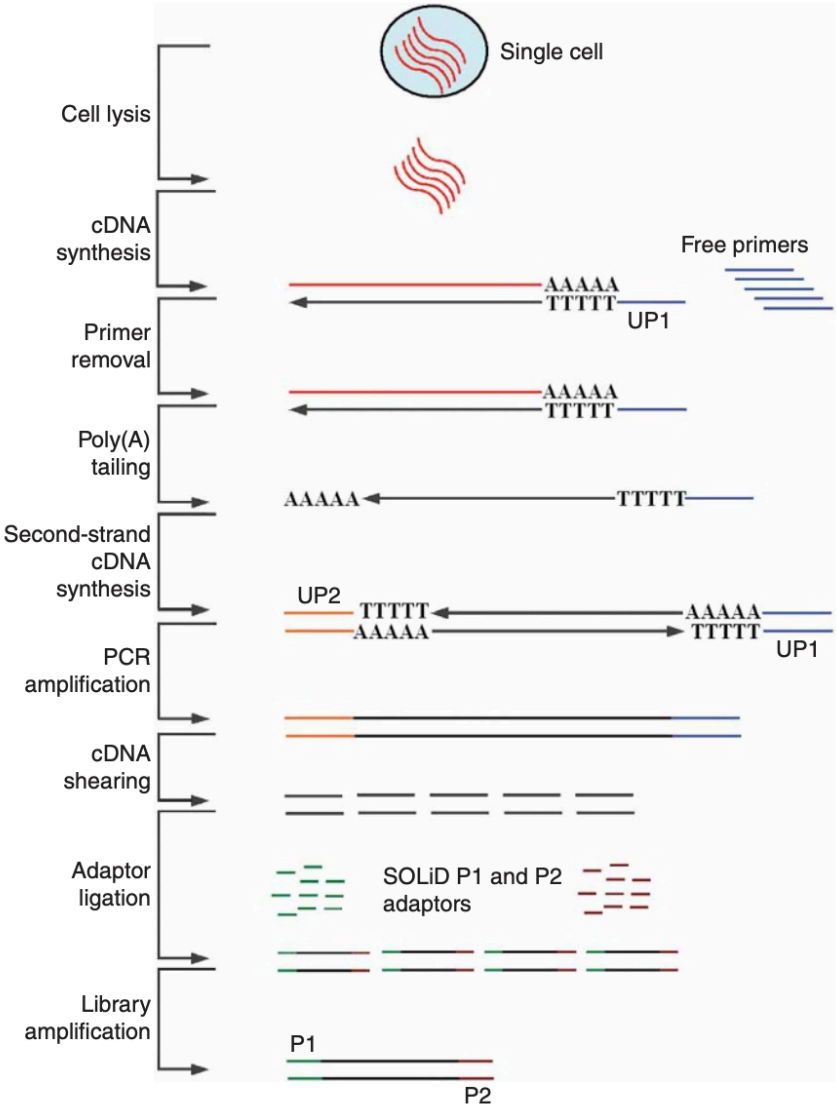


Figure 1.8: Schematic of single cell RNA sequencing as described by Tang et al., 2009.
A single cell is manually picked under a microscope and lysed. Then mRNAs are reverse-transcribed into cDNAs using a poly(T) primer with anchor sequence (UP1) and unused primers are digested. Poly(A) tails are added to the first-strand cDNAs at the 3′ end, and second-strand cDNAs are synthesized using poly(T) primers with another anchor sequence (UP2). Then cDNAs are evenly amplified by PCR using UP1 and UP2 primers, fragmented, and P1 and P2 adaptors are ligated to the ends. Finally, emulsion PCR is performed by mixing libraries with 1 μm diameter beads with P1 primers covalently attached to their surfaces.

Microfluidic technology is another promising method for isolation of single cells (Whitesides, 2006). This technology has gained wide popularity due to low sample volume required and low handling cost. A widely used commercial platform, Fluidigm C1, is based on microfluidic

technology. The platform provides automated single-cell lysis, RNA extraction, and cDNA synthesis for up to 800 cells in parallel on a single chip and offers lower false positives and less bias than tube-based technologies. However, its major drawbacks include the number of cells (>1000) required for capture and the homogeneous size limit of the cells being analyzed. Microdroplet-based microfluidics circumvent some of these challenges (Utada *et al.*, 2005). The technology allows monodispersion of aqueous droplets in a continuous oil phase. The lower volume required by this system compared to standard microfluidic chambers enables the manipulation and screening of thousands to millions of cells at a reduced cost. The commercial chromium system from 10X genomics based on this technology, offers high-throughput profiling of 3′ ends of RNAs of single cells with high capture efficiency, further enabling the analysis of rare cell types in a sufficiently heterogeneous biological space. Another methodology to isolate circulating tumour cells (CTCs) has been developed by CellSearch. The technology uses a magnet conjugated with antibodies to detect CTCs of epithelial origin in patient blood samples.

Table 1.1: Summary of current scRNA-seq methods (Modified from (Hedlund and Deng, 2017))

Methods	Captured RNAs	cDNA coverage	Amplification technology
Homopolymer tailing based-PCR			
Tang 2009	polyA+ RNAs	Full-length with 3' biased	PCR after poly A tailing
Quartz-seq	polyA+ RNAs	Full-length with 3' biased	PCR after poly A tailing
SC3-seq	polyA+ RNAs	Full-length with 3' biased	PCR after poly A tailing
SUPeR-seq	polyA+ RNAs and polyA- RNAs	Full-length	PCR after poly-dA tailing
MATQ-seq	polyA+ RNAs and polyA- RNAs	Full-length	PCR after poly-dC tailing

Switching mechanism at 5' end of RNA Template (SMART)-based PCR			
STRT-seq	polyA+ RNAs	5'Tag(TSS)	Template switching-based PCR
Smart-seq	polyA+ RNAs	Full length with weak 3'-biased	Template switching-based PCR
Smart-seq2	polyA+ RNAs	Nearly full-length	Template switching-based PCR
Drop-seq	polyA+ RNAs	3' tag (UTR)	Template switching-based PCR
Patch-seq	polyA+ RNAs	Nearly full-length 5' tag (TSS)	Template switching-based PCR
in vitro transcription-based linear amplification			
CEL-seq	polyA+ RNAs	3' tag (UTR)	*in vitro* transcription
CEL-seq2	polyA+ RNAs	3' tag (UTR)	*in vitro* transcription and random primer based PCR
MARS-seq	polyA+ RNAs	3' tag (UTR)	*in vitro* transcription
inDrops	polyA+ RNAs	3' tag (UTR)	*in vitro* transcription
Designed primers-based PCR			
DP-seq	polyA+ RNAs	Specific gene-region	PCR using designed heptamers
Cyto-seq	polyA+ RNAs	3' tag (UTR)	gene-specific primer based PCR
MALBAC RNA	polyA+ RNAs	Full-length	Quasilinear PCR with 7 random MALBAC primers

The generation of scRNA-seq libraries require cell lysis, reverse transcription into first-strand cDNA, second-strand synthesis, and cDNA amplification. Generally, the cells are lysed in a hypotonic buffer, and poly(A)+ selection is performed using poly(dT) primers to capture mRNAs. Due to the selection step based on poly(dT) primers, non-polyadenylated RNA species, including mostly non-coding RNAs and circular RNAs are precluded. It has been well

established that due to Poisson sampling, only 10–20% of transcripts will be reverse transcribed at this stage (Islam *et al.*, 2014a). This low mRNA capture efficiency is an important challenge that remains in existing scRNA-seq protocols and necessitates a highly efficient cell lysing strategy.

Next step in the process of scRNA-seq is preparation of cDNA. The first strand synthesis for cDNA preparation is done by an engineered version of the reverse transcriptase isolated from Moloney murine leukaemia virus (M-MLV) having low RNase H activity and increased thermostability (Arezi and Hogrefe, 2009). Once the first strand synthesis is done, second strand synthesis can be achieved using either poly(A) tailing (Sasagawa *et al.*, 2013) or by a template-switching mechanism also known as SMART technology (switching mechanism at 5′ end of RNA template) (Islam *et al.*, 2011). The SMART technology approach ensures uniform coverage without loss of strand-specificity compared to the former. Some of the scRNA-seq methods which follow SMART technology include Smart-seq (Switch Mechanism at the 5' End of RNA Templates sequencing), Smart-seq2 and STRT (Single-cell tagged reverse transcription sequencing).

Another approach ligates the 5’ end of cDNA with either poly(A) or poly(C) to build common adaptors for PCR amplification. However, the exponential amplification generated by PCR can potentially skew the representation of gene expression profiles towards shorter and less G-C rich amplicons. To avoid these drawbacks, CEL-seq (Cell Expression by Linear Amplification Sequencing) method was devised (Hashimshony *et al.*, 2012), which is the first method based on *in vitro* transcription (IVT). Although this method allows linear amplification of templates, it is quite time consuming, as it requires an additional reverse transcription, which may lead to 3′ coverage biases (Morris, Singh and Eberwine, 2011). In addition to this, another method for cDNA amplification is based on Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) which uses quasilinear instead of exponential amplification to amplify and sequence the transcriptomes of single cells (Zong *et al.*, 2012). The MALBAC primers generate amplicons to have complementary ends that loop and prevent DNA from being copied exponentially, thereby avoiding amplification bias.

Next step in scRNA-seq is the preparation of sequencing library. As mentioned in Table 1.1, there are several methods which have been developed for scRNA-seq. These methods can be generally divided into two types- full-length and tag-based. Full-length methods are advantageous in achieving a uniform gene body read coverage and increase the number of mappable reads. Such broad coverage is required for the analyses of the data for isoform discovery, splicing events and SNP identification for allelic gene expression. However, one of the major drawbacks of full-length methods is that it was not possible to multiplex and pool all samples for one tube preparation of Illumina sequencing libraries, which increased cost and labour. Most scRNA-seq methods are tag-based and can be further divided into two categories depending on if they utilize a 3′ (e.g CEL-seq/CEL-seq2, MARS-seq) or a 5′ tag (e.g STRT). One of the advantages of tag-based methods is that these can be combined with Unique Molecule Identifiers (UMIs), which is a sequence of several nucleotides which can be integrated into each transcript by reverse transcription (Kivioja *et al.*, 2012). This enables multiplexing of more samples and improvement of gene-level quantification and throughput. So far, tag-based methods have relatively low sensitivity as mappable reads are restricted to one end of the transcript. Thus, tag-based methods are mostly used for gene expression quantification and cannot be utilized for isoform identification or splicing.

Due to stochasticity in transcription and the scarcity of RNA material in an individual cell, scRNA-seq data often face challenges in determining cell-to-cell transcriptional variability. While some of this variability may be inherent due to biological factors such as stochastic gene expression, different cell cycle state, cell size etc., the observed variability may also be attributed to technical factors such as RNA capture efficiency, random dropouts during library preparation or varying sequencing depth. Further, experimental batch effects may be introduced during sample handling steps, capturing of single cells, sequencing at various sequencing depths using different lots of experiments involving several biological specimens (Leek *et al.*, 2010). One way to identify the presence of batch effects, is to visualize the data using principal component analysis (PCA) and observe if cells are grouped by their experimental origin. Furthermore, multiple batch correction methods have been established to deal with these issues (Johnson, Li and Rabinovic, 2007). Addition of spike-in RNA standards of known abundance to the endogenous samples has also become a common practice to better account for technical variability due to random dropout events during library preparation

(Baker *et al.*, 2005). However, the ideal concentration of spike-in RNA standards is also cell-type specific and may not be the same for control and perturbation samples.

Despite the above discussed challenges in analyzing scRNA-seq data, the technology has been successfully implemented in understanding cell-to-cell transcriptional variability during the development of drug tolerance in cancer cells (K. T. Kim *et al.*, 2015) and in reconstructing clonal and phylogenetic relationships, during the process of cancer evolution, by modelling transcriptional kinetics (Müller *et al.*, 2016). In addition to these, scRNA-seq has been recently proven to have potential in performing lineage tracing during evolution by reconstructing lineage phylogeny over many generations (Frieda *et al.*, 2017). As sequencing costs decrease, it will be possible to routinely analyze more than a million cells within the next 5 years (Svensson, Vento-Tormo and Teichmann, 2018). The Human Cell Atlas aims to map 35 trillion cells from the human body and then use gene expression profiles to classify and identify new cell types (Regev *et al.*, 2017). Recent progress in overcoming the challenges in different single cell profiling datasets have allowed the integration of scRNA-seq with epigenomic profiling such as scATAC-seq (Mezger *et al.*, 2018) which has enhanced our understanding of correlating transcriptional variability with differential chromatin accessibility patterns at single-cell level (Stuart *et al.*, 2019). Another example of recent development in the field of single cell analysis is the idea of an elegant methodology named “spatial transcriptomics”. In this method, the transcriptome can be analyzed in intact tissues sections on slides, without the need for cell isolation (Ståhl *et al.*, 2016).

Future advancements may involve combined profiling of genome, epigenome, transcriptome and/or protein from the same single cell. This will enable a comprehensive understanding of each cell by integrating information from DNA, RNA, protein as well as epigenetic modifications and would enable identification of a rare subset of population contributing towards disease evolution. Few examples of such methods that have already been developed to achieve this include G&T-seq (genome and transcriptome sequencing) for genomic and transcriptomics analyses (Macaulay *et al.*, 2015), scTrio-seq (single-cell triple omics sequencing) for genomic, transcriptomic analyses together with DNA methylome (Hou *et al.*, 2016), scM&T-seq (single-cell genome-wide methylome and transcriptome sequencing) for transcriptomics and methylome (Angermueller *et al.*, 2016). Studies analyzing genome and

transcriptome simultaneously can correlate chromosomal CNVs (copy-number variations), chromosome fusion and SNVs (single nucleotide variations) in regulatory elements with the corresponding gene expression level. Such studies can also reveal clonal structure and cell subtypes within the population, which directly links genotype variation with phenotype outcomes (Macaulay *et al.*, 2015). On the other hand, incorporation of transcriptomic and methylomic analysis would reveal how the degree of epigenetic variability in the form of DNA methylation of different functional elements in the genome may reflect cell-to-cell gene expression variability in single cells during disease development.

1.5 Acute Myeloid Leukaemia- a cancer model to study the dependency on non-genetic variability

Acute Myeloid Leukaemia (AML) is a heterogeneous clonal disorder of haematopoietic progenitor cells ('blasts') which lose the ability to differentiate normally and to respond to normal regulators of proliferation (Lowenberg, Downing and Burnett, 1999). The abnormal differentiation of myeloid cells results in enhanced production of immature malignant cells with fewer differentiated red blood cells (RBCs), white blood cells (WBCs) and platelets. AML is more prevalent in adults who are 65 years or old with a dismal prognosis of <30% 5-year survival (National Cancer Institute). A better understanding of how AML evolves using recently developed next-generation sequencing technologies, can help us devise strategies to improve the therapy and prognosis of AML patients.

Haematopoietic Stem Cells (HSCs), like other stem cells, are undifferentiated long-lived cells capable of asymmetric division, facilitating both self-renewal and the generation of differentiated progeny. Additionally, these cells can undergo either self-renewing (clonal expansion) or differentiating (clonal extinction) symmetric division (Pina and Enver, 2007). On an average, human HSCs undergo cell division once every 40 weeks (Catlin *et al.*, 2011). However, blood cell production is a continuous process throughout life, with an adult human producing an estimated 10^{11} cells daily (Beerman *et al.*, 2010). These properties make HSCs, like other tissue stem cells, prime targets for malignant transformation. On the other hand, the fact that some mutations can transform differentiating cells as well suggests that HSCs might not be the only target for disease transformation (Huntly *et al.*, 2004).

1.6 Genetic, Epigenetic and Transcriptional variability in Acute Myeloid Leukaemia

As discussed previously that somatic mutations get accumulated during the course of disease development and are the source of genetic variability amongst a given population of cells over a period of time. It is worth noting that AML has one of the lowest number of mutational load per case of any adult cancer studied to date where the range varies widely between individual cases (Lawrence *et al.*, 2013b). However, the number of mutations in an individual HSC increases almost linearly with age and is very similar to that found in *de novo* AML, suggesting that AML develops stochastically in a cell, which fortuitously accrues a transforming combination of mutations (Welch *et al.*, 2012c).

AML has relatively well-defined set of recurrent mutations, most of which fall into functional categories (TCGA, 2013b; Papaemmanuil *et al.*, 2016). A study conducted by The Cancer Genome Atlas (TCGA) in 2013 on 200 AML patient samples suggested the presence of at least one recurrent mutation in 199 (>99%) samples (TCGA, 2013b). These mutations were found to be present in one of the nine categories that were defined according to biological function and which have a putative role in AML pathogenesis: transcription-factor fusions (18% of cases), the gene encoding nucleophosmin (NPM1) (27%), tumour-suppressor genes (16%), DNA-methylation–related genes (44%), activated signalling genes (59%), chromatin-modifying genes (30%), myeloid transcription-factor genes (22%), cohesin-complex genes (13%), and spliceosome-complex genes (14%) (TCGA, 2013b) (Fig 1.9).

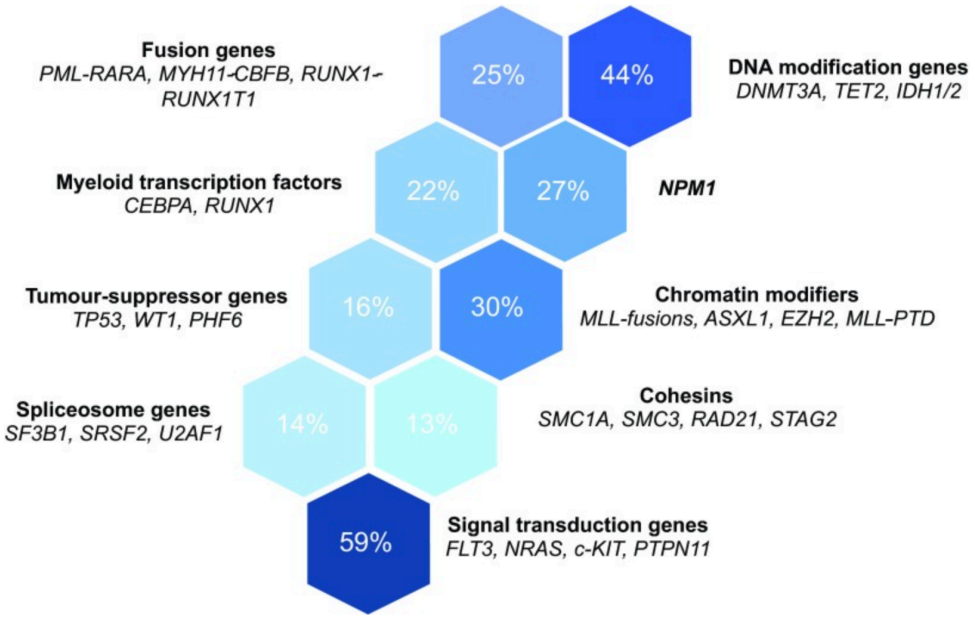


Figure 1.9: Genes recurrently mutated in AML belong to distinct functional groups or pathways. The most prominent functional groups and genes associated with these are listed. The proportion of AMLs with mutations affecting each of these groups is displayed. Adapted from (Grove and Vassiliou, 2014), based on the data obtained from (TCGA, 2013b).

The somatic mutations in AML can be divided into two categories based on their functions. The first class of mutations mainly enhance proliferation and survival of haematopoietic progenitors through the activation of signalling pathways. Few examples include Fms-Related Tyrosine Kinase 3 (FLT3), C-Kit, and Neuroblastoma RAS Viral (V-Ras) Oncogene Homolog (NRAS). On the other hand, the second class of mutations mainly occur in transcription factors such as RUNX1, CEBPα and RARA, where these mutations hinder cell differentiation and create an accumulation of immature progenitors. Based on the frequent observation of mutations in each of these two classes in AML, Gilliland & Griffin proposed a 'double-hit' model of leukaemogenesis (Kelly, Clark and Gilliland, 2002). They proposed that either class of mutations is insufficient to cause malignant transformation of haematopoietic stem cells on its own. Only the co-occurrence of a class I and class II mutations would cooperate to form leukaemia. However, there were certain limitations to the double-hit model, where not every AML harboured the mutations that strictly correspond to these two classes. This was in-line with a recent epigenomics study which specifically showed that the combination of Tet2 loss with expression of $FLT3^{ITD}$ results in synergistic and gain-of-function effects on the epigenetic state and on transcription. Overall, the study showed that combination of a class I and class II mutations, rather than causing additive effects, instead cooperated in a synergistic manner to

induce distinct epigenetic and transcriptional programming effects than what was caused by either mutation alone (Shih *et al.*, 2015).

AML has been observed as a disease model with a complex mosaic of cells containing different combinations of genetic lesions and epigenetic variants (Ding *et al.*, 2012; Li *et al.*, 2014; Paguirigan *et al.*, 2015). The clonal evolutionary dynamics in AML is marked by continuous acquisition and loss of specific mutations, sometimes occurring at different timepoints which lead to simultaneous evolutionary convergence and divergence among particular clones and subclones during the course of disease (Ding *et al.*, 2012; Paguirigan *et al.*, 2015; Wu *et al.*, 2020). Exposure to chemotherapy regimens places an enormous stress on AML cell populations and may be more toxic to certain clones than others. Remarkably, it has been observed that clonal composition of AMLs can change quite markedly after therapy in relapsed disease, with selection occurring at both the genetic and epigenetic levels (Ding *et al.*, 2012; Li *et al.*, 2014). These observations highlighted the functional significance of the genetic composition of clones. Certain clones within the same AML patient have been characterized by distinct morphology, differentiation markers, and engraftment potential in immunocompromised mice (Klco *et al.*, 2014). These characteristics of AML cells suggest the fact that the subpopulation of more immature leukaemia stem cells may have the potential to perpetuate and repopulate the disease, and these cells are more chemotherapy resistant. Hence leukaemia stem cell functionality may be a reflection of subclonal combinations of specific mutations, with the emergence of chemoresistance and/or tumour changes potentially driven by these clonality shifts (Li *et al.*, 2014).

Apart from these clonal shifts, the variability in the order of acquisition of mutations also contributes towards AML pathogenesis. Somatic mutations in the epigenetic modifiers DNMT3A, IDH1/2 and TET2 are initiating mutations in AML, especially those with normal karyotype (Welch *et al.*, 2012c; Shlush *et al.*, 2014a; Xie *et al.*, 2014). Mutations of the NPM1 gene that encode the aberrant NPM1c protein have also been described as initial hits (Welch *et al.*, 2012c). However, clonal evolution analysis of AML patients at relapse suggests that, in the AML cases having both DNMT3A and NPM1, the DNMT3A mutations may precede NPM1 (Krönke *et al.*, 2013) since DNMT3A mutation persisted at relapse in cases where NPM1 mutation was lost. Another study conducted by Shlush and colleagues also established a

sequential order of mutation acquisition whereby DNMT3A and IDH2 mutations exist in pre-leukaemic HSCs and precede NPM1c and FLT3-ITD (Shlush *et al.*, 2014a). In-line with this, Corces-Zimmerman *et al.* also found that somatic mutations in epigenetic modifiers which regulate cytosine methylation occur early in pre-leukaemia cells and persist once AML goes in remission, whereas somatic mutations in signalling pathways that drive proliferation are later events in AML transformation (Corces-Zimmerman *et al.*, 2014a). These data suggest that disruption of epigenetic patterning is likely an early and prominent event during leukaemogenesis. On similar lines, somatic mutations in transcription regulators including NPM1, FLT3, IDH1, N/KRAS, RUNX1, CEBPA, WT1, PTPN11, c-KIT are observed quite often. For example, NPM1 occurs in 27% of *de novo* AML cases. Likewise, chromosomal translocations (often a fusion of a transcription factor gene with another gene) occur frequently in *de novo* AML (18%). Examples include fusion of genes such as RUNX1 with ETO, PML with RARA, CBFB with MYHII, as well as fusion of MLL with various partner genes (TCGA, 2013b; Koeffler and Leong, 2017). These findings overall indicate that early disruption in transcriptional landscape is a common hallmark in AML. This was compatible with the findings on transcriptional variability suggested by Pina *et al.* in haematopoietic system (Fig 1.10A) and Buganim *et al.* in induced Pluripotent Stem Cells (iPSCs) obtained from fibroblasts (Fig 1.10B), where the authors had shown that early committed progenitors show higher variability in their transcriptional programmes compared to the multipotent self-renewing cells (Pina *et al.*, 2012) (Buganim *et al.*, 2012). These findings altogether strongly suggest that epigenetic and transcriptional reprogramming are key mechanisms during the early stage of AML development and may play a significant role in contributing towards cancer cell fitness.

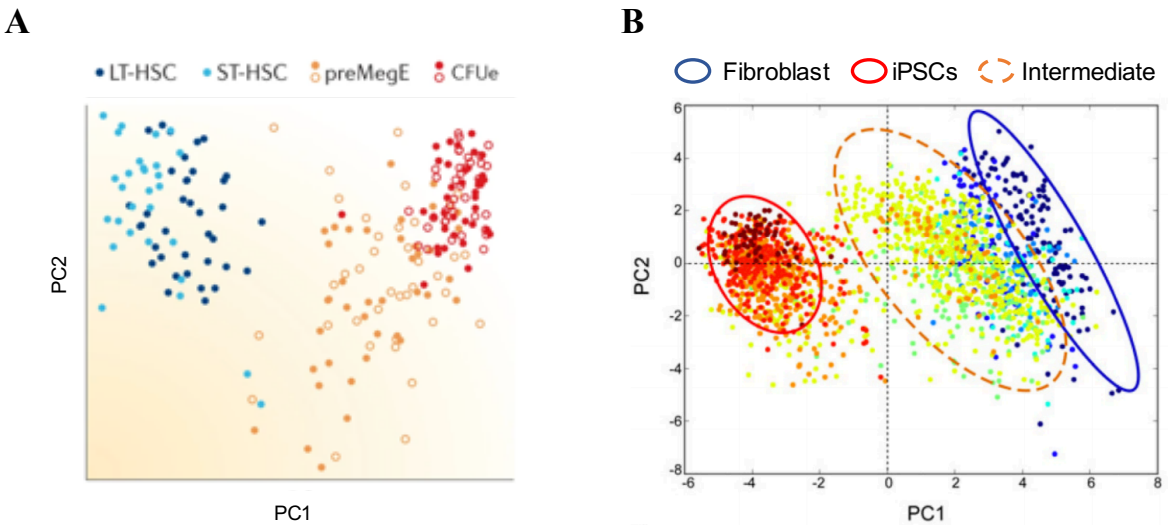


Figure 1.10: Enhanced transcriptional variability during intermediate stage of cellular differentiation process.

(A) Principal component analysis (PCA) plots of single-cell quantitative reverse transcription PCR data for cultured, primary mouse bone marrow (BM) cells in two different cytokine conditions, undergoing erythroid commitment and differentiation decisions (data obtained from (Pina *et al.*, 2012)). The plot highlights the point that early committed cells, pre-megakaryocytic/erythroid progenitors (preMegE), are more heterogeneous in their transcriptional programmes than the multipotent self-renewing cells including long-term reconstituting haematopoietic stem cells (LT-HSCs) and short-term HSCs (ST-HSC) they originate from, or the differentiated progeny, colony forming unit- erythroid (CFUe) they give rise to. Adapted from (Moris, Pina and Arias, 2016a), **(B)** PC projections of individual cells, coloured by their sample identification. The blue circle surrounds fibroblasts population and the red circle surrounds iPSCs. The orange dotted circle surrounds a third intermediate population with enhanced transcriptional variability. Modified from (Buganim *et al.*, 2012).

Based on the above inferences that epigenetic and transcriptional variability may act at early stages of AML progression, I focussed on the AML mouse models, which require cooperating mutations for leukaemia progression (Fig 1.11) and thus have prolonged ‘pre-leukaemic’ state.

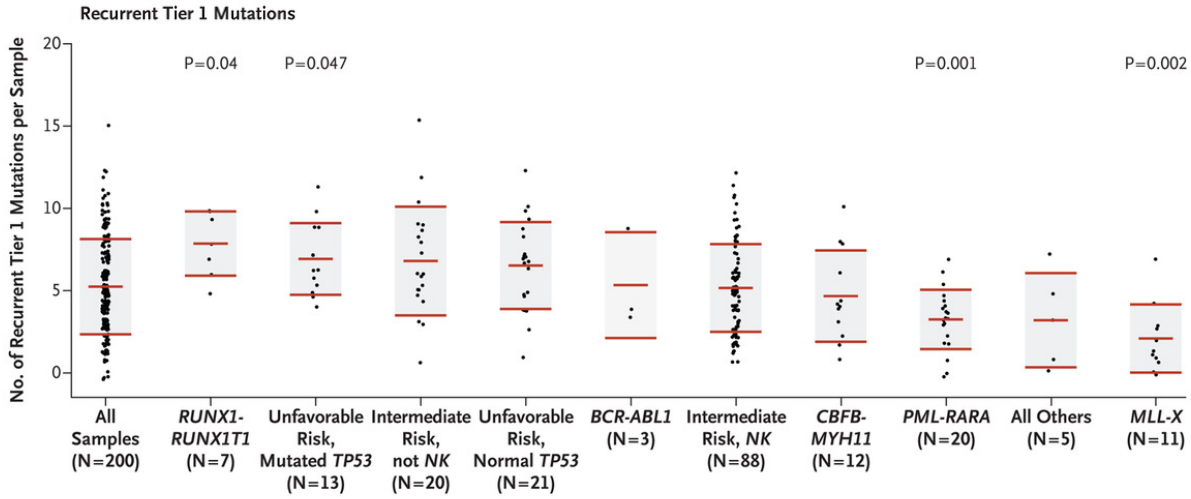


Figure 1.11: Recurrent mutations in different cytogenetic and mutational backgrounds in AML. Recurrent tier 1 mutations in 200 AML patient samples. For each set of data, the middle horizontal line indicates the mean, and the shaded area indicates ±1 SD. P values are shown for the groups that had significant differences from the mean number of recurrent tier 1 mutations in all samples. NK denotes normal karyotype. Adapted from (TCGA, 2013b).

1.7 Pre-Leukaemia

Pre-Leukaemia or pre-leukaemic phase is defined as the sequential acquisition of mutations in functionally normal HSCs which give rise to pre-leukaemic HSCs. These pre-leukaemic HSCs are capable of increased cell proliferation or self-renewal but retain normal multilineage potential and possess a high risk of leukaemic transformation (P.Greenberg, 1983; Jacobs, 1985). The term 'pre-leukaemia' was first coined by Block et al. in 1953, where it was defined as heterogeneous group of haematopoietic disorders associated with a block in myeloid differentiation and chronic cytopenias (Block, Jacobson and Bethard, 1953). Pre-leukaemia was initially reported in concordant twins with ALL, whose haematopoietic cells shared a unique somatic rearrangement involving the MLL gene (Ford *et al.*, 1993). The current model of pre-leukaemic HSCs has been defined on the basis of various scientific evidence which suggest that first leukaemogenic mutation must either occur in a cell that is capable of self-renewal or confer self-renewal upon the cell. If the first mutation fails to meet one of these two criteria, it will be lost over time due to terminal differentiation (Weissman, 2005).

The term pre-leukaemia has undergone a major evolution since its inception by Block et al. in 1953. The term also refers to the heterogeneous group of haematopoietic disorders associated with a block in myeloid differentiation and chronic cytopenias, known as myelodysplastic syndromes (MDS) (Haferlach *et al.*, 2014). As the population ages, the prevalence of these abnormalities markedly increases, where about 75 new patients per 100,000 individuals per year over the age of 65 years develop MDS. About 20–30% of these individuals do indeed progress to AML; and thus, these individuals might justifiably be known as preleukaemic (Koeffler and Leong, 2017).

Another scenario where the existence of pre-leukaemic clones has been reported is the X-inactivation during blastogenesis which usually results in equal inactivation of the maternal and paternal X chromosomes in the female (Gale *et al.*, 1993). Few studies have indicated that older individuals have haematopoietic cells that favour usage of one X allele. This has been explained by unequal exhaustion of haematopoietic stem cells as individuals age. This observation was buttressed by genomically marking murine haematopoietic stem cells and using them for complete haematopoietic transplantation. Multiple clones of haematopoiesis initially developed in these transplanted animals; but as the mice aged, oligoclonal haematopoiesis predominated, suggesting that with age, certain haematopoietic clones can predominate (Williams *et al.*, 1984).

The presence of inherited germline mutations of one of several transcription factors (RUNX1, GATA2 or CEBPA) in certain individuals has been shown to have an increased propensity to develop AML over their life (Smith *et al.*, 2004; Wlodarski *et al.*, 2016). Other inherited germline mutations that can also predispose to MDS include Diamond–Blackfan anaemia (20% risk of MDS/AML), dyskeratosis congenita (disorder of telomere maintenance, 30% risk MDS/AML), Fanconi anaemia (40% risk), severe congenital neutropenia (20–40% risk), Shwachman-Diamond syndrome (10–35% risk), mutations of the SRP72 gene (increased risk of aplastic anaemia or MDS), Bloom syndrome (25% risk for MDS/AML), Li-Fraumeni syndrome (p53 mutation) (5–7% risk) and familial monosomy (Babushok and Bessler, 2015). For each of these inherited germline mutations, their progression to frank AML requires the clonal acquisition of additional genetic mutations (Churpek *et al.*, 2015).

The presence of pre-leukaemic clones has also been reported in certain AML patients who achieved complete morphologic remission by chemotherapy and continued to have an abnormal clone (Fialkow, Janssen and Bartram, 1991). This observation suggested that preleukaemic clone can be used to describe the loss of the major phenotype of AML, including driver mutations associated with AML blast cells, but where the clonal haematopoiesis persisted, and could evolve into relapse AML (Koeffler and Leong, 2017). Ley and colleagues had reported that ~50% of the individuals whose AML cells carried DNMT3A, TET2 or IDH1/2 mutations at diagnosis, continued to have the same mutation in >5% of the haematopoietic cells at complete morphologic remission, but the classical driver mutations (for example, NPM1, FLT3 and RAS) found at diagnosis, were not detectable at remission (Klco *et al.*, 2015). Another study performed by Dick's group on a cohort of individuals whose AML blast cells had mutations of DNMT3A and NPM1 at both diagnosis and relapse, indicated the presence of only DNMT3A mutation upon morphologic complete remission (Shlush *et al.*, 2014b).

In conclusion, pre-leukaemia might be defined as a condition that has modifying mutations in the bone marrow that either cause MDS or cause clonal haematopoietic expansion initially without disease but associated with progression to AML (Koeffler and Leong, 2017).

During the course of my PhD, I focussed on two different mouse models of pre-leukaemia-

1.7.1 RUNX1-RUNX1T1(9a) model

RUNX1-RUNX1T1 (aka AML1-ETO) is characterized by chromosomal translocation, t(8;21)(q22;q22), and is one of the most common genetic abnormalities in AML, identified in 15% of the cases in AML (Miyoshi *et al.*, 1991). Since the discovery of pre-leukaemia clones in monozygotic twins with ALL, the clonotypic RUNX1–RUNX1T1 fusion sequences were also detected in Guthrie spots in cases of childhood AML (Wiemels *et al.*, 2002). The prevalence of detectable RUNX1–RUNX1T1 in cord blood is 100-fold greater than the risk of the corresponding leukaemia, and the frequency of positive cells (10^{-4} to 10^{-3}), indicating substantial clonal expansion of the abnormal progenitor population of cells (Mori *et al.*, 2002). One of the reasons for this observation is the fact that these fusion genes are not sufficient for

disease development, as indicated by protracted post-natal latencies, non-concordant phenotypes in monozygotic twins (Wiemels *et al.*, 2002) and the lack of overt leukaemia in transgenic mice (Rhoades *et al.*, 2000). Therefore, secondary genetic events appear necessary for tumour development.

The RUNX1-RUNX1T1 fusion with breaks at 8q22 and 21q22.3 was first reported by Dr Janet Rowley in 1973 during the analysis of a leukaemia patient sample (Rowley, 1973). Both the RUNX1 and ETO genes were identified subsequently by several groups in the early 1990s (Gao *et al.*, 1991; Miyoshi *et al.*, 1991). RUNX1 (also known as AML1, CBFα2 or PEBP2αB) gene belongs to the family of Runt-related transcription factors (RUNXs) (Van Wijnen *et al.*, 2004). It is known as acute myeloid leukaemia 1 (AML1) as its gene sequence was discovered from a human patient with AML (Miyoshi *et al.*, 1991). The expression of RUNX1 marks the earliest haematopoietic precursor cells (T North, 1999). Mice haploinsufficient for RUNX1 displayed a reduced number of HSCs defined by Lineage$^-$ Sca1$^+$ Ckit$^+$ (LSK) (Sun and Downing, 2004). RUNX1 gene is found to be highly mutated in AML patients and is the most common target of chromosomal rearrangements in AML (Osato *et al.*, 1999). RUNX1T1 gene (also known as ETO or MTG8), on the other hand, is mainly characterized by its four Nervy homology regions (NHRs), which are homologous to Drosophila nervy protein (Feinstein *et al.*, 1995). These NHRs define the domains of ETO gene that mediate interactions with other proteins. NHR1 share high similarity with TATA-binding protein associated factor (TAF) proteins and inhibits the ability for E proteins to recruit co-activator molecules like p300/CBP (Eriksson, Lennartsson and Lehmann, 2015)(Peterson and Zhang, 2004). NHR2 can mediate homo- and hetero- oligomerization of ETO and can interact with co-repressors like Histone deacetylases (HDACs) (Liu *et al.*, 2006). NHR3 shares structural homology with Protein Kinase A (PKA) and binds to PKA regulatory subunit (PKA RIIalpha) (Fukuyama *et al.*, 2001). NHR4 contains a zinc chelating motif and interacts with co-repressors NCoR/SMRT (Liu *et al.*, 2007).

The RUNX1-RUNX1T1 fusion occurs in intron 5 of the RUNX1 locus and in either intron 1a or 1b of the ETO locus (Tighe and Calabi, 1995; Tighe, Daga and Calabi, 1995; Zhang *et al.*, 2002) (Fig 1.12). The fusion protein consists of 752 amino acids involving N-terminal portion of RUNX1, including its DNA binding domain, and almost the entire RUNX1T1 protein

(Miyoshi *et al.*, 1993) (Fig 1.13A). Although the presence of RUNX1-RUNX1T1 leads to alteration of gene expression and haematopoietic cell proliferation, its expression does not lead to development of leukaemia (Peterson and Zhang, 2004). This is indicative of the weakly oncogenic effect of RUNX1-RUNX1T1, which require cooperating mutations in order to develop leukaemia.

Figure 1.12: Genomic structure of t(8;21).
The translocation occurs in intron 5 of *AML1* and intron 1a or 1b of *ETO* gene. Modified from (Yan *et al.*, 2006).

An alternatively spliced isoform of the RUNX1-RUNX1T1 transcript, RUNX1-RUNX1T1 (9a) was identified (Yan *et al.*, 2006) which includes an extra exon, exon 9a, of the RUNX1T1 gene. RUNX1-RUNX1T1(9a) encodes a C-terminally truncated RUNX1-RUNX1T1 protein of 575 amino acids (Fig 1.13B). Expression of RUNX1-RUNX1T1(9a) leads to rapid development of leukaemia in a mouse retroviral transduction–transplantation model (Yan *et al.*, 2006).

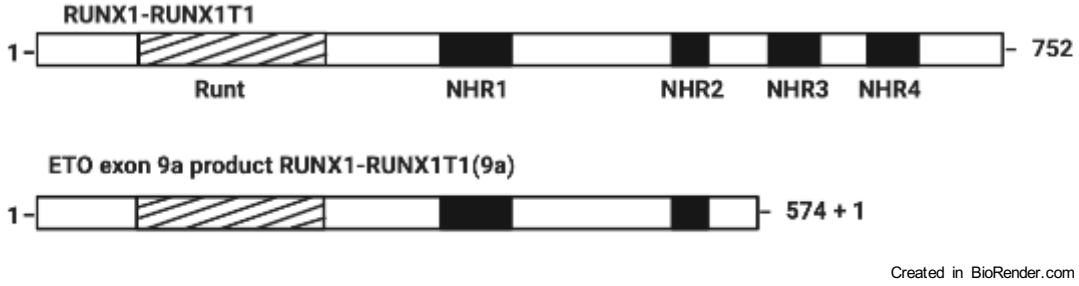


Figure 1.13: Structure of the full-length RUNX1-RUNX1T1 protein and truncated protein.
(A) Structure of full length RUNX1-RUNX1T1 encoding for 752 amino acids**, (B)** Structure of truncated RUNX1-RUNX1T1 with ETO exon 9a product encoding for 575 amino acids. The number after the "+" sign indicates the number of extra amino acids that were not included in the original RUNX1-RUNX1T1 sequence. Modified from (Yan *et al.*, 2006).

Subsequent to the identification of the spliced form of RUNX1-RUNX1T1, namely RUNX1-RUNX1T1(9a), Chen and colleagues explored the presence of mutations in *C-KIT* gene and expression in t(8;21) leukaemogenesis (Wang *et al.*, 2005; Jiao *et al.*, 2009). They found that *C-KIT* was highly expressed in 81.3% of patients with t(8;21) compared to patients with other leukaemias. Further, among patients with t(8;21) and mutated *C-KIT*, most of the leukaemic cells at disease presentation harboured both of the genetic alterations, whereas in three such cases investigated during complete remission, only t(8;21), but not mutated *C-KIT*, could be detected, suggesting that mutated *C-KIT* should be a subsequent event on the basis of t(8;21).

The epigenomics studies performed on deciphering the molecular pathogenesis of RUNX1-RUNX1T1 suggested depletion of RUNX1-RUNX1T1 affects transcriptional programmes associated with myeloid differentiation, proliferation and self-renewal, in addition to those promoting cell cycle progression and DNA synthesis (Ptasinska *et al.*, 2012). RUNX1-RUNX1T1 cells were found to require active RUNX1 where RUNX1 knockdown in Kasumi-1 cells resulted in apoptosis (Schnittger *et al.*, 2011). Another study demonstrated that RUNX1-RUNX1T1 transformed cells require a functional RUNX1 to maintain adequate PU.1 levels, which is critical for RUNX1-RUNX1T1 leukaemia development in mouse transplantation models (Ben-Ami *et al.*, 2013; Nafria *et al.*, 2020). Increasing evidence suggests that the early presence of RUNX1-RUNX1T1 may promote mutagenesis. Several studies showed that RUNX1-RUNX1T1 downregulates genes involved in base-excision repair mechanism (BER) as well as in homologous recombination (HR) (Krejci *et al.*, 2008; Forster *et al.*, 2016) suggesting that RUNX1-RUNX1T1 may compromise genome integrity by impairing DNA stress monitoring and DNA repair capabilities. RUNX1-RUNX1T1 has been also found to be capable of recruiting the histone acetyltransferase (HAT) p300/CBP complex. RUNX1-RUNX1T1 changes the local nucleosomal environment by acetylating histone lysine residues and recruiting the transcriptional machinery thus playing an essential role in gene regulation in normal haematopoiesis (Kasper *et al.*, 2002; Rebel *et al.*, 2002). Another recent study identified the role of RUNX1-RUNX1T1 in contributing to high order cis regulatory interactions by making use of Promoter Capture Hi-C (CHi-C) method in Kasumi-1 cells. The depletion of RUNX1-RUNX1T1 led to a rewiring of promoter-enhancer interactions, which was driven by increased C/EBPα and loss of AP-1 binding after knockdown (Ptasinska *et al.*, 2019). Overall, these studies highlighted the molecular mechanisms impacted by RUNX1-RUNX1T1 during

leukaemia progression and maintenance, however, the detailed molecular mechanism underlying stepwise RUNX1-RUNX1T1(9a) progression still remain unexplored and further studies are needed to understand the RUNX1-RUNX1T1(9a) pathogenesis.

1.7.2 Idh1R132H model

Isocitrate Dehydrogenase 1 (IDH1) is a key metabolic enzyme and belongs to one of the three isoforms of IDH enzymes present in eukaryotes. IDH1 and IDH2 are homodimeric and dependent on Nicotinamide adenine dinucleotide phosphate ($NADP^+$), while IDH3 is a structurally distinct heterotetrameric enzyme that utilizes Nicotinamide adenine dinucleotide (NAD^+) as a co-factor (Fig 1.14). The cytoplasmic IDH1 contributes to the metabolic regulation of cells by catalyzing the oxidative decarboxylation of isocitrate to produce carbon dioxide (CO_2) and α-ketoglutarate (α-KG) in a $NADP^+$ dependent manner (Yang *et al.*, 2012a). Whereas the IDH1 present in peroxisomes is responsible for providing reducing equivalents to support lipid biosynthesis and redox homeostasis (Jo *et al.*, 2001; Lee *et al.*, 2002; Kim *et al.*, 2007).

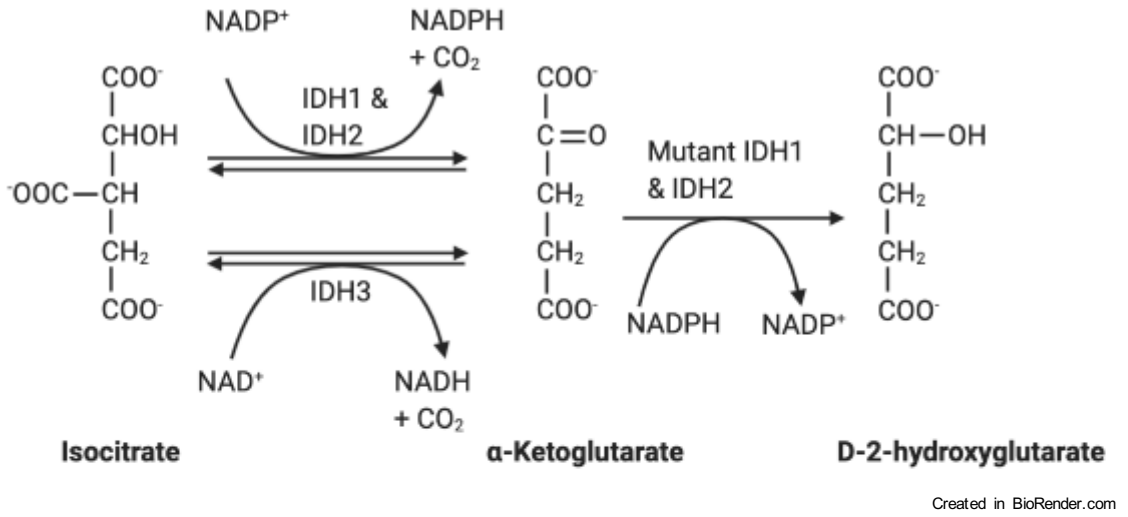


Figure 1.14: Chemical reactions catalyzed by the wild-type IDH enzymes and tumour-derived IDH1/2 mutants.

The only structural difference between α-KG and D-2-HG is the replacement of the 2-ketone group in α-KG by a hydroxyl group in 2-HG. ICT- Isocitrate, α-KG- α-ketoglutarate and D-2-HG- 2-Hydroxyglutarate. Adapted from (Yang *et al.*, 2012b).

IDH1/2 mutations were first identified in colorectal cancer, and later, were also identified in brain tumours, with >70% incidence in secondary gliomas (Yan *et al.*, 2009). Whole-genome sequencing analysis (WGS) of AML patients indicated the presence of IDH1/2 mutations in 12-18% of AML cases (Mardis *et al.*, 2009). Following the discovery of IDH1/2 mutations in glioma and AML, these were found in several different cancers, including thyroid cancer, intrahepatic cholangiocarcinoma, cartilaginous tumour, and some other cancers (Mi *et al.*, 2009; Hemerly, Bastos and Cerutti, 2010; Amary *et al.*, 2011; Wang *et al.*, 2013). Recently, IDH1/2 mutations have been identified as an important regulator in promoting cardiac arrest as well (Kattih *et al.*, 2020). This studies altogether suggested the significant relevance of these mutations in disease and in particular, in cancer development.

The most common IDH1 mutations are found on a single amino acid residue, Arg 132 (R132) which are heterozygous (Yan *et al.*, 2009) and believed to have a dominant-negative effect (Zhao *et al.*, 2009). IDH1 R132 mutant gains a neomorphic activity and convert α-KG to D-2-Hydroxyglutarate (D-2-HG or 2-HG), (also known as R(-)-2-hydroxyglutarate) by utilizing NADPH, a key component of cellular anti-oxidation systems (Dang *et al.*, 2009). 2-HG is maintained at very low levels in normal cells and tissues and is not known to have any significant role in any metabolic pathway. The cancer cells harbouring IDH1 mutant have up to ~100-fold greater 2-HG levels as compared to wild type hence, termed as an 'oncometabolite' (Dang *et al.*, 2009).

2-HG and α-KG share highly identical structure. As α-KG apart from playing a crucial role in Krebs cycle, is also involved in other biochemical pathways, including glutamate synthesis, transamination of amino acids, generation of NADPH and a cofactor for dioxygenase enzymes, the structural similarity of 2-HG and α-KG may lead to competitive inhibition of several other α-KG dependent dioxygenases (Xu *et al.*, 2011). Some of these dioxygenases include Ten-eleven translocation (TET) family, Jumonji-C domain-containing histone demethylases, cytochrome C oxidase and Hypoxia inducible factor 1 α (HIF1α) (Figueroa *et al.*, 2010; Xu *et al.*, 2011). Since these enzymes are responsible for epigenetic modifications in normal cells, alteration in these in cancer cells due to accumulation of 2-HG may constitute a key hallmark of tumourigenesis (Figueroa *et al.*, 2010).

IDH1/2 mutations have been identified as recurrent mutations in case of normal cytogenetic AML (NC-AML) which may lead to leukaemogenesis (Mardis *et al.*, 2009). These mutations were found to have increased frequency with increasing age (Fathi *et al.*, 2015) where mutant NPM1 and FLT3 were identified as their potential collaborators. However, the presence of IDH1/2 mutations mutually exclusively of the transcription-factor fusions, suggests that these mutations may have functions in the initiation of AML that are similar to the functions of fusion genes (TCGA, 2013b) (Fig 1.15). IDH1 mutations most often involve a cysteine (R132C) or histidine (R132H) substitution for arginine at R132. To date, the only *in vivo* study conducted by Sasaki et al. on *Idh1* R132H mutant knock-in mouse model suggested the development of haematologic abnormalities including anaemia, splenomegaly and extramedullary haematopoiesis. The proportion of progenitor population of cells characterized by Lineage$^-$ Sca1$^+$ Kit$^+$, in the BM was expanded by about 5-fold in 42-46 week old mice, but the proportion of common myeloid progenitors (CMPs), granulocyte monocyte progenitors (GMPs), megakaryocyte erythrocyte progenitors (MEPs), and common lymphoid progenitors (CLPs) in the bone marrow remained unaffected (Sasaki *et al.*, 2012a). Progression to acute leukaemia was not reported, indicating that additional collaborating mutations are required in-line with the model utilised in this thesis work.

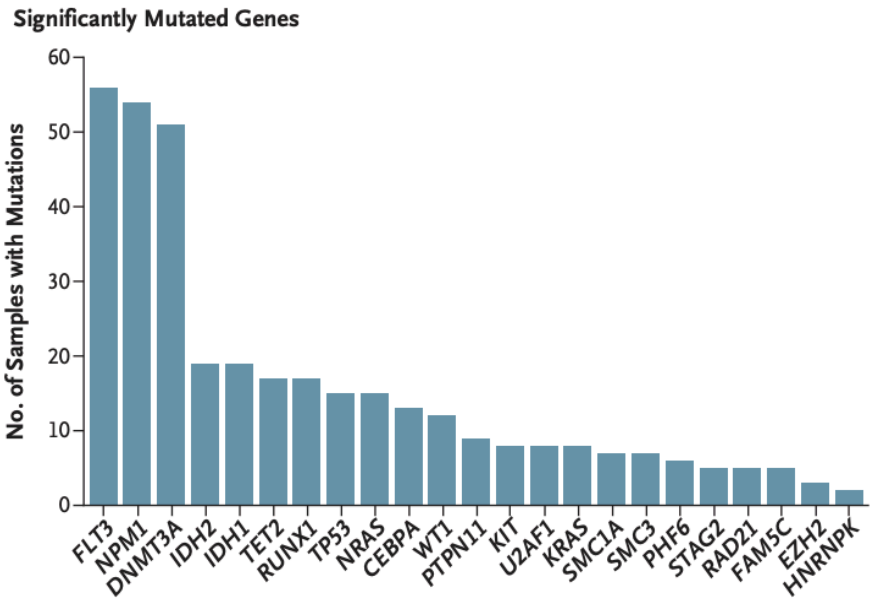


Figure 1.15: TCGA analysis on recurrent mutations in 200 AML patient samples.
The plot shows significantly mutated genes and the number of samples with each mutation. Adapted from (TCGA, 2013b).

Further studies have also found that these mutations are acquired early in the progression from normal haematopoietic stem/progenitor cells to form leukaemia (Corces-Zimmerman *et al.*, 2014b; Shlush *et al.*, 2014c) and are stable during the evolution (Chou *et al.*, 2010), indicating that a population of IDH1 mutant cells survive initial chemotherapy and contribute to relapse. Due to the recent discovery of recurrent IDH1 mutations in AML and unavailability of sufficient data on understanding the disease progression, it is quite important to understand the stepwise progression of this disease with the contribution of genetic and non-genetic factors during the course of disease evolution.

1.8 MLL-AF9 model- a maintenance model of leukaemia

Mixed lineage leukaemia gene (MLL or KMT2A) is located on 11q23 which is commonly involved in rearrangements (Djabali *et al.*, 1992). MLL rearrangements fuse the N-terminus of MLL to a fusion partner protein and constitute >70% of infant ALL and between 35-50% of infant AML; MLL translocations also occur in leukaemias of older children and adults, overall accounting for 10% of cases (Krivtsov and Armstrong, 2007). Generally, patients with MLL rearrangements have a poor prognosis and are treated according to high-risk protocols, depending on the translocation partner fused to MLL gene (Meyer *et al.*, 2009). In a genomic study conducted by TCGA in 2013, the group of patients with MLL fusions had the fewest recurrent tier 1 mutations, suggesting that MLL fusions require fewer cooperating mutations than other AML-initiating events (TCGA, 2013b) (MLL-X in Fig 1.11).

MLL is expressed in most tissues, including myeloid and lymphoid cells, and positively regulates expression of the clustered Hox genes through histone H3 lysine 4 (H3K4) methyltransferase activity (Milne *et al.*, 2002). Knockout studies in mice show that Mll is important for the maintenance of Hox gene expression. The expression of Hox genes is initiated normally in Mll knockout mice but is not properly maintained (Yu *et al.*, 1995). Further analysis of the role of Mll in haematopoietic system revealed that Mll is necessary for the proliferation and/or survival of the HSC and progenitor compartment in both the developing foetus and adult mice (Yu *et al.*, 1998).

ALL1-fused gene from chromosome 9 (AF9) is one of the most common fusion partners of MLL1, and this fusion arise from the t(9;11)(p22,q23) translocation (Meyer *et al.*, 2013). The MLL-AF9 fusion protein consequential to t(9;11) arrangement is found in about 2-5% of all AML and up to 25% of *de novo* AML in children (Huret *et al.*, 2000). MLL-AF9 induced leukaemia has an aggressive phenotype with median survival for *de novo* cases being only ~4 years (Huret *et al.*, 2000). The first *MLL–AF9* fusion knock-in mouse model was made in 1996 by Corral and colleagues. The model constitutively expressed *MLL–AF9* under the control of the endogenous MLL promoter and developed AML (Corral *et al.*, 1996). Subsequently, another conditional knock-in model was developed that showed *MLL-AF9* expression in long-term haematopoietic stem cells (LT-HSC) and *in vitro* resulted in dispersed clonogenic growth and expression of genes involved in migration and invasion (Stavropoulou *et al.*, 2016)

Delving deeper into the mechanistic understanding of MLL-AF9 pathogenesis, Somervaille and colleagues had shown that the strong oncogenic activity of MLL-AF9 cells aided in immortalization of colony-forming cells and displayed high leukaemia maintenance potential (Somervaille and Cleary, 2006). This was confirmed based on mice transplantation experiment using *in vitro MLL-AF9* transformed colonies maintained in a semi-solid medium. All of the animal recipients injected with *MLL-AF9* transformed cells developed AML with a median latency of 84.5 days demonstrating the fact that strong oncogenicity of *MLL-AF9* transformed cells could aid in maintaining the leukaemia initiating cells (Somervaille and Cleary, 2006). These observations were further confirmed by another study conducted by Horton *et al*., where the authors found that the immortalization of leukaemogenic population of cells was dependent on constitutive MLL-AF9 expression (Horton *et al.*, 2013).

These observations of immortalization of AML transformed cells in the presence of strong oncogenic effect of MLL-AF9 fusion protein, makes it an excellent model to study AML maintenance in contrast to the pre-leukaemia models, namely *RUNX1-RUNX1T1(9a)* and *Idh1*R132H, as discussed above.

1.9 Ribosomal Biogenesis

The eukaryotic ribosome is a complex macromolecular machine made of 4 rRNA species and 80 ribosomal proteins (RPs) (Warner, 1999). The mature ribosome is composed of 2 subunits, the small 40S ribosomal subunit containing the 18S rRNA and 33 RPs and the large 60S ribosomal subunit containing the 28S, 5.8S, and 5S rRNAs and 47 RPs. Ribosome biogenesis, one of the most complex and energy consuming process in the cell (Warner, 1999; Ben-Shem *et al.*, 2011), involves the coordinated work of the three RNA polymerases (RNA pol) and the assistance of more than 200 protein co-factors (Henras *et al.*, 2008). This stepwise journey commences in the nucleolus with the synthesis by RNA pol I of a large RNA transcript, the 45S pre-rRNA, which encodes three of the mature rRNA species (18S, 5.8S, and 28S rRNAs). The 5S rRNA is transcribed in the nucleoplasm by RNA Pol III and imported to the nucleolus. These rRNAs are then engaged in a series of modifications comprising nucleolytic processing steps and successive recruitment of RPs (transcribed by RNA pol II) in order to shape precursor ribosomal particles. Then, following further nucleoplasmic maturation steps, pre-60S and pre-40S ribosomes are eventually translocated to the cytoplasm where they undergo final maturation steps to form the mature 40S and 60S ribosomal subunits and achieve translation competence (Tschochner and Hurt, 2003). Ribosome biogenesis is directly regulated by the extracellular environment and stresses, nutrient availability, and cell proliferation and has been extensively studied in yeast as well as in higher eukaryotes, both in normal and pathological contexts (Bastide and David, 2018; Klinge and Woolford, 2019).

In mammalian cells, RP expression varies greatly among tissues and even from one cell type to another (Bortoluzzi *et al.*, 2002; Kondrashov *et al.*, 2011a; Signer *et al.*, 2014). This suggests that RPs expression pattern may evolve along cell differentiation, concomitantly with the acquisition of specialized functions. This is exemplified by RPL38 tissue-specific expression pattern during murine embryogenesis, which correlates with tissues that are affected by the loss of function of this protein (Kondrashov *et al.*, 2011a). A recent study performed on similar lines also showed that the protein synthesis is dynamically controlled in a cell-type- and the developmental stage-specific manner in the haematopoietic system (Magee and Signer, 2021). Adult HSCs depend on maintaining low protein synthesis (Signer *et al.*, 2014, 2016), and modest increases in protein synthesis partially impair adult HSC self-renewal by increasing the

biogenesis of misfolded proteins (Hidalgo San Jose *et al.*, 2020). Whereas fetal HSCs maintain relatively high protein synthesis, yet they retain self-renewal capacity, suggesting that they can cope with misfolded proteins more effectively than adult HSCs (Magee and Signer, 2021).

Ribosome biogenesis and protein synthesis are involved in a considerable number of components and events and are highly regulated in order to efficiently respond to extrinsic demands. In cancer cells, disruption of ribosome biogenesis and protein synthesis is associated with altered expression of key genes encoding translation initiation factors and proto-oncogenes such as mTOR, c-MYC, and RAS (Ruggero *et al.*, 2004; Rosen and She, 2006; Markman, Dienstmann and Tabernero, 2010). One of the most striking examples connecting RP haploinsufficiency with elevated cancer incidence are ribosomopathies, a collection of disorders, in which genetic abnormalities trigger impaired ribosome biogenesis and function (De Keersmaecker, Sulima and Dinman, 2015). These syndromes result in specific clinical phenotypes that can be categorized as cellular hypoproliferative defects, often involving bone marrow failure and/or craniofacial or other skeletal defects. Remarkably, some of these diseases are associated with increased cancer risk, although the type and frequency vary significantly (De Keersmaecker, Sulima and Dinman, 2015). The best-studied ribosomopathy is named Diamond-Blackfan anemia (DBA) and characterized by bone marrow failure syndrome with a severe erythroid defect. Although this disease was first associated with recurrent mutations in RPS19 gene, further studies identified mutations or deletions in other RPs (Sulima *et al.*, 2017). DBA patients show a predisposition, particularly, towards Myelodysplastic syndrome (MDS) or AML development (Vlachos *et al.*, 2012a). Other congenital disorders have been associated with defective ribosome biogenesis and cancer predisposition, including Schwachman–Diamond syndrome, Dyskeratosis Congenital (DC), cartilage hair hypoplasia, and Treacher Collins syndrome (Narla and Ebert, 2010). In the 5q syndrome, a subtype of adult myelodysplastic syndrome, the long arm of chromosome 5 is deleted, resulting in RPS14 haploinsufficiency and subsequent severe refractory anaemia (Mills and Green, 2017).

Quite often, mutation or alteration of a single ribosomal component triggers global changes of ribosome-related genes expression. Although such changes may reflect a mere necessity to sustain high protein synthesis rate and rapid cell proliferation, they may hold prognostic or

predictive values. For example, translational profiling analysis of chronic lymphocytic leukaemia (CLL) patients led to the identification of a ribosome-related translational signature comprising RPs, translation initiation factors, and DKC1. The authors showed that decreased DKC1 expression observed in CLL was associated with reduced synthesis of certain RPs, which in turn promoted translatome alteration and the acquisition of an aggressive phenotype (Sbarrato *et al.*, 2016).

Ribosome alteration can arise following somatic mutation of RP gene or as an indirect consequence of changes in ribosome-related genes' expression. Recently, genome-wide analysis of tumour samples with next generation sequencing technologies revealed frequent somatic defects in multiple RP genes. In T-cell acute lymphoblastic leukaemia (T-ALL), somatic mutations and deletions of RP encoding genes have been reported in about 20% cases, the most frequent ones on RPL10 (8% of pediatric T-ALL cases) and RPL22 (10%), with rare defects in RPL5 (2%) and RPL11 (1.4%) (De Keersmaecker *et al.*, 2013a; Tzoneva *et al.*, 2013a). In aggressive CLL, somatic missense mutations of RPS15 occur in 10–20% of patients (Landau *et al.*, 2015; Ljungström *et al.*, 2016a).

Ribosomal alteration may also trigger translational dysfunction, which will have an impact on gene expression. For instance, the T-ALL-associated RPL10-R98S mutation triggers profound structural, biochemical, and translational fidelity defects that may drive cancer evolution through gene expression reprogramming (Ljungström *et al.*, 2016a). Given its role in driving cancer evolution, it would be worth understanding whether attenuation in ribosomal biosynthetic programmes also alter gene expression landscape during pre-leukaemia transformation.

1.10 Lysine acetyltransferase 2a *Kat2a*/KAT2A- a tool to study non-genetic variability

Lysine Acetyltransferase 2a or Kat2a is a histone acetyl transferase (HAT) which belongs to Gcn5-related N-acetyltransferase (GNAT) superfamily (Neuwald and Landsman, 1997). Kat2a is one of the mammalian homologous (other one is Kat2b) orthologues of yeast General control nonderepressible 5 (Gcn5) which was the first histone specific HAT to be isolated from Tetrahymena and was the first enzyme to link histone acetylation and transcriptional activation

(Brownell *et al.*, 1996). Kat2a is mostly abundant in haematopoiesis, neural tissue and has a role in development (Yamauchi *et al.*, 2000). It plays an important role in developing embryo where it is ubiquitously expressed between E7.5 and E9.0, however, its expression decreases after E16.5 indicating the minimal requirement of Kat2a during terminal differentiation (Xu *et al.*, 2000a). Kat2a null mice die at E10.5, with extensive mesodermal apoptosis; suggesting that Kat2a is important during embryonic development (Xu *et al.*, 2000b). Kat2a is also required in pluripotent stem cells system not from the maintenance perspective but for the stabilization of pluripotency gene regulatory networks, as our lab has shown previously (Moris *et al.*, 2018a).

There is literature on associating the role of Kat2a in cancer, however, the mechanistic link of Kat2a during tumour progression still needs to be unravelled. The expression of KAT2A is found to be associated with poor prognosis in breast cancer (Chen *et al.*, 2010), non-small cell lung carcinoma (N-SCLC) (Chen *et al.*, 2013), colon cancer (Yin *et al.*, 2015) and renal cell carcinoma (Hong *et al.*, 2020). The poor prognosis is majorly associated with histone acetylation-mediated co-activation of E2F and MYC transcriptional targets to maintain cell proliferation and survival (Yin *et al.*, 2015). Increased expression of KAT2A also associates with poor survival in renal cell carcinoma, on contrary to this, overexpression of KAT2A in pancreatic adenocarcinoma and glioma provides a survival advantage to the cells (Arede and Pina, 2020). A recent study has linked the KAT2A succinylation activity to maintenance of pancreatic adenocarcinoma cell lines, overall raising the possibility of the contrasting role of KAT2A in a cancer stage-specific manner, as shown for other epigenetic regulators (Basheer *et al.*, 2019). It is worth noting that till date no recurrent mutations in KAT2A have been described in any cancer type suggesting hijacking of KAT2A at an epigenetic level for the establishment or maintenance of cancer.

Kat2a forms a crucial part of Spt-Ada-Gcn5 acetyltransferase (SAGA) which is a multifunctional co-activator with distinct activities including HAT, a histone deubiquitinase (DUB), and an activator-binding module (Lee *et al.*, 2011). Kat2a was identified not only in SAGA, but also as a subunit of a second co-activator complex with HAT activity in Drosophila and mammals, named Ada Two A Containing (ATAC) (Guelman *et al.*, 2006). The mammalian SAGA complex is composed of 20 subunits organised in 4 distinct functional and

structural modules namely, the HAT module, central core of SAGA including SPT and TAF module and DUB module (Fig 1.16). The HAT module which is also shared with ATAC consists of KAT2A (or KAT2B, mutually exclusive to each other), SGF29, TADA3 and a complex-specific ADA2 variant –ADA2B in SAGA and ADA2A in ATAC. SGF29 is responsible for recruiting KAT2A on active promoters marked by H3K4 tri or di-methylation (Bian *et al.*, 2011). TADA3 is a transcriptional activator adaptor required for transcriptional activity (Martinez *et al.*, 2001) whereas ADA2B is a zinc finger protein which is capable of binding double-stranded DNA (Zhang *et al.*, 2019). TAF5L and TAF6L present in the central core of SAGA and are specific to SAGA, whilst the other TAFs are shared with TFIID complex, the RNA Polymerase II (Pol II) General Transcription Factor. TRRAP is a large transcription factor interaction module present in SAGA complex and was originally identified as cofactor of c-Myc and E2F proteins (McMahon *et al.*, 1998). Finally, the DUB module present in SAGA complex is catalyzed by USP22, which targets H2B and H2A as well as non-histone proteins (Armour *et al.*, 2013). All 4 members of the DUB module – USP22, ATXN7L3, ATXN7 and ENY2 – are needed for full deubiquitinating activity (Köhler *et al.*, 2008).

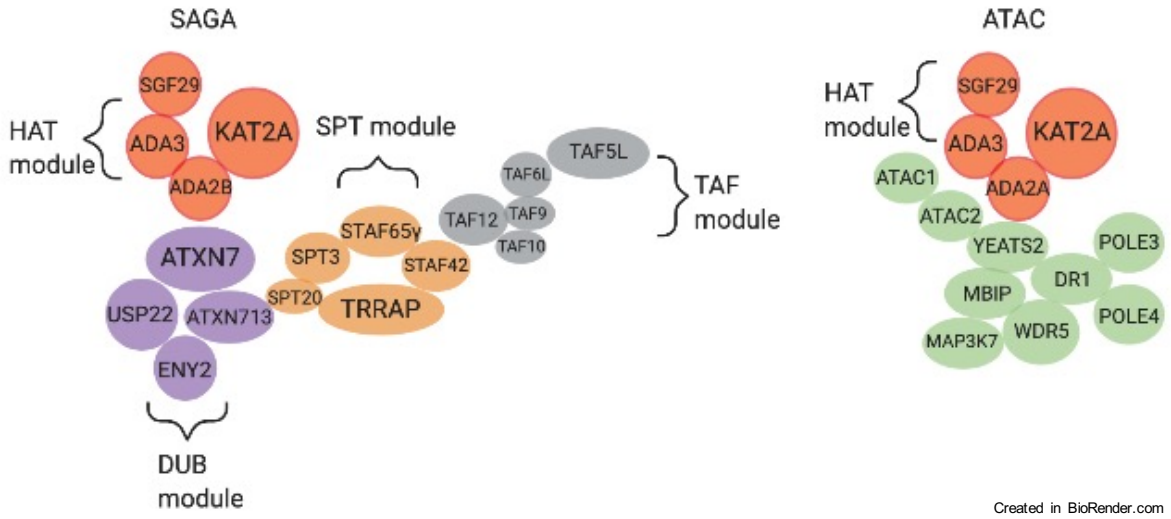


Figure 1.16: Schematic of KAT2A-containing complexes and their functions and implications in diseases.
This figure only summarizes the presence of subunits, not their arrangement within ATAC. The modules are indicated by the following colour codes: red, HAT module; blue, DUB module; orange, SPT module; grey, TAF module; and green, ATAC-specific unit. DUB: Deubiquitination; HAT: Histone acetyltransferase; TF: Transcription factor. Modified from (Wang and Dent, 2014).

KAT2A is present as a subunit in another complex known as ATAC which includes YEATS2, DR1 (NC2β), ATAC1 (ZZZ3), WDR5, MBIP and ATAC2 (KAT14) within the ATAC subunit (Fig 1.16). YEATS2 reads H3K27 acetylation mark which allows for integration of additional transcription activation signals (Mi *et al.*, 2017). DR1 is a member of NC2 complex, which heterodimerizes with YEATS2 to interact with TBP. ATAC1 is a zinc finger protein that specifically binds H3 tails (W. Mi *et al.*, 2018) and WDR5 which has been identified as a candidate therapeutic target in CEBPA N-terminal leukaemia (Grebien *et al.*, 2015). MBIP is a MAP3K regulator present only in mammalian cell complexes (Wang *et al.*, 2008) whereas ATAC2 is responsible for a second HAT activity in this complex (Guelman *et al.*, 2009). YEATS2, MBIP and KAT14 are required for complex integrity (Guelman *et al.*, 2009).

The yeast orthologue of Kat2a, known as Gcn5, is a classical regulator of transcriptional variability (Raser and O'Shea, 2004b) as it's loss promotes cell-to-cell transcriptional variability (Weinberger *et al.*, 2012a). As discussed above, transcriptional variability is inherent to the phenomenon known as transcriptional bursting which depends on various factors including variability in TATA-box, nucleosome positioning. Cell-to-cell transcriptional variability is one of the non-genetic mechanisms which lead to imbalance between self-renewal and cellular differentiation, thereby may contribute towards cancer evolution.

Delving deeper into understanding the role of *KAT2A*, our lab has recently shed some light on its role in AML. Using a Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-dropout screen of AML cell lines, KAT2A was identified as a genetic vulnerability in AML (Tzelepis *et al.*, 2016). Our lab has shown that the acetyltransferase activity of KAT2A is responsible for maintaining undifferentiated cultured and patient-derived human AML cells *in vitro*. Further, our lab studied the *MLL-AF9* leukaemia using a conditional *Kat2a* knockout mice model (Domingues *et al.*, 2020) where we observed that loss of *Kat2a* impact the long-

term preservation of functional leukaemia stem-like cells (LSC). Further, the *Kat2a*-depleted LSCs lost repopulating capacity and did not fully progress through myelo-monocytic differentiation. Upon performing scRNA-seq of *Kat2a* wild-type (WT) and knockout (KO) leukaemia, an increase in cell-to-cell transcriptional variability was observed in KO, supporting a role of *KAT2A* in transcriptional stability. Additionally, upon learning the differentiation trajectories from leukaemia stem cells to differentiated leukaemia cells, the single-cell trajectory analysis performed on WT and KO cells highlighted an almost linear trajectory for WT *MLL-AF9* cells, whereas *Kat2a* KO leukaemia cells were distributed along multiple discontinuous differentiation trajectories (Fig 1.17). Overall, this suggested that *Kat2a* KO cells had multiple uncoordinated routes into cell fate decision-making, which were initiated but not coherently completed by cells depleted of stem cell potential. These set of observations were interesting as they suggest cell-to-cell transcriptional variability as a mechanism where the leukaemia transformed cells lose the capability of maintaining a balance between self-renewal and differentiation, thereby setting up a paradigm of utilising *Kat2a* as a tool to understand cell-fate decisions in a given disease model.

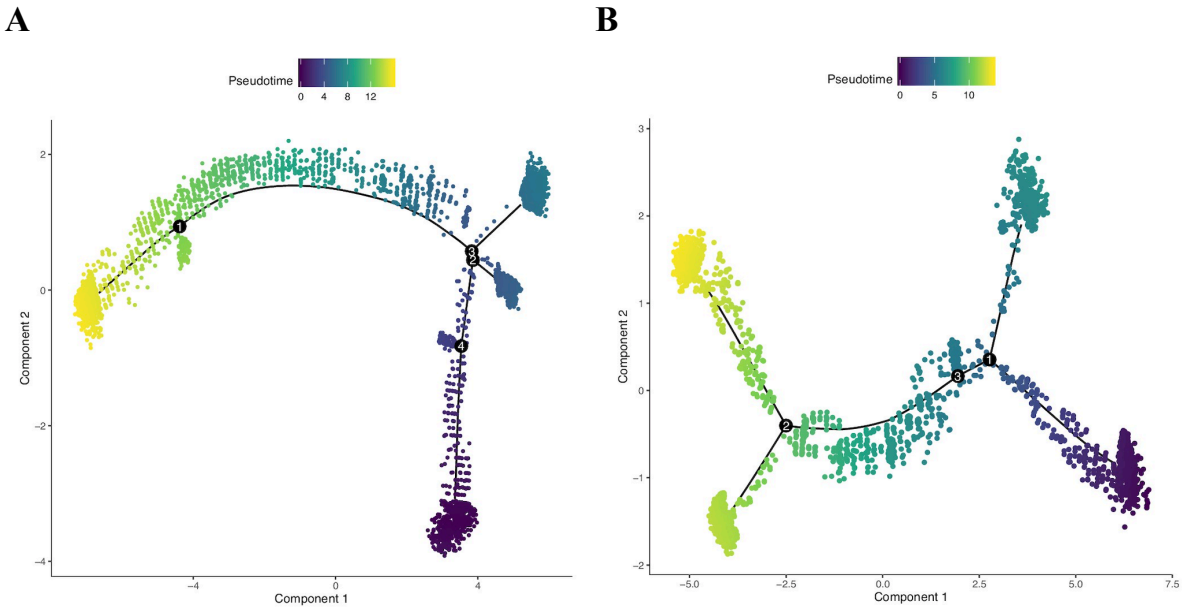


Figure 1.17: *Kat2a* WT and *Kat2a* knockout *MLL-AF9* primary leukaemias have unique differentiation trajectories.

(A) Monocle pseudotime trajectory of *MLL-AF9* transformed *Kat2a* WT cells**, (B)** Monocle pseudotime trajectory of *MLL-AF9* transformed *Kat2a* NULL cells. Adapted from Domingues and colleagues (Domingues *et al.*, 2020).

1.11 Hypothesis and Rationale

The central role of non-genetic mechanisms, in particular cell-to-cell transcriptional variability in impacting the cell-fate decision process during disease progression independent of genetic blueprint is evident from the literature. The fact that Acute Myeloid Leukaemia does not possess high mutational burden and has a strong dependency on epigenetic and transcription factors for disease progression makes it an excellent model to study the contributions of non-genetic variability during leukaemia propagation. The Pina lab has previously shown that KAT2A/*Kat2a* inhibition enhances transcriptional variability in mouse embryonic stem cells (Moris *et al.*, 2018a), a finding that is recapitulated in the *MLL-AF9* leukaemia model (Domingues *et al.*, 2020). While transcriptional variability may be disadvantageous to MLL-AF9 transformed cells, which are abnormally retained in leukaemia self-renewal through the activity of the fusion protein, I propose that variability in gene expression may equally promote transition from pre-leukaemia towards leukaemia just as achieved by accumulation of mutational events. To test this hypothesis, I have made use of *RUNX1-RUNX1T1(9a)* and *Idh1*R132H mouse models of pre-leukaemia, which are normally reliant on additional mutational hits for full leukaemia transformation. Specifically, I analysed the biology of *RUNX1-RUNX1T1(9a)*-initiated and *Idh1*R132H-initiated AML in a conditional *Kat2a* knockout background. In order to study the potential link between loss of *Kat2a*, its consequent increase in gene expression variability, and pre-leukaemia progression, I performed single-cell RNA sequencing (scRNA-seq) of early-stage *Kat2a* WT and *Kat2a* NULL *RUNX1-RUNX1T1(9a)* pre-leukaemia samples. The scRNA-seq analysis was utilized to study the transcriptional programmes impaired upon loss of *Kat2a* during pre-leukaemia transformation which were further validated experimentally and contrasted with MLL-AF9 model of leukaemia. Further, to study the lineage dependencies and cell fate decisions during pre-leukaemia progression upon loss of *Kat2a,* pseudo-temporal ordering analyses were performed. Overall, *Kat2a* depletion portrays an interesting model of perturbation of AML progression that can either prevent or accelerate leukaemia progression depending on mutational context which can be further extended to study other pre-malignancies.

1.12 Objectives

In order to test the mechanistic contributions of cell-to-cell transcriptional variability consequent to *Kat2a* loss during leukaemia initiation, I aimed to answer the following questions-

1. Does *Kat2a* play a functional role in leukaemia initiation? Specifically, is leukaemia initiation impacted in *RUNX1-RUNX1T1(9a)* (aka *AML1-ETO9a*) and *Idh1*R132H mice models of leukaemia with a *Kat2a* conditional knockout background?

2. What are the transcriptional programmes impacted upon loss of *Kat2a*? Do these programmes coincide with the general process of pre-leukaemia transformation?

3. How do these transcriptional programmes affect pre-leukaemia transformation in *RUNX1-RUNX1T1(9a)* and *Idh1*R132H pre-leukaemia models in contrast to *MLL-AF9* leukaemia?

4. What is the role of these transcriptional programmes in carrying out cellular variability during the process of pre-leukaemia progression? Is this cell-to-cell transcriptional variability associated with *Kat2a* loss a consequence of epigenetic variability?

1.13 Thesis Structure

Following the literature review in the 'Introduction' chapter, the thesis work begins with Chapter-2 which describes the experimental design and methodology of the experiments, which are further discussed in detail in the results section. The methodology section includes descriptions of the *Kat2a* conditional knockout mouse model as well as the setting up and validation of *Idh1*R132H colony in a conditional *Kat2a* knockout background. The computational pipelines and algorithms followed for single-cell RNA sequencing analysis are also described in this section.

The results from the analysis and experiments are then divided amongst the next four chapters. Chapter-3 describes the functional characterization of *RUNX1-RUNX1T1(9a)* and *Idh1*R132H pre-leukaemia in a *Kat2a* conditional genetic background. The functional analysis of *RUNX1-RUNX1T1(9a)* model suggested an accelerated leukaemia progression upon loss of *Kat2a*. The pre-leukaemia analysis further showed perpetuation of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* knockout cells with an enhanced self-renewal capacity *in vitro*. Similar analysis was performed on *Idh1*R132H model where progression to leukaemia was not observed, suggesting requirement of additional incorporating events. However, an increase in progenitor cell population was observed, consistent with *RUNX1-RUNX1T1(9a)* model. The pre-leukaemia analysis highlighted an increase in self-renewal capacity *in vitro* during early stages of pre-leukaemia progression.

In Chapter-4, I discuss the single-cell RNA sequencing analysis of *RUNX1-RUNX1T1(9a)* time-series pre-leukaemia in order to study the potential link between loss of *Kat2a* and the consequent increase in gene expression variability. The chapter first describes the pre-processing steps performed, which include filtering of poor-quality cells and genes and normalization to exclude bias introduced by variable sequencing depth. The chapter also describes dimensionality reduction analysis performed using principal component analysis and t-distributed stochastic neighbour embedding to define a robust gene set for downstream analysis. I then discuss the transcriptional programmes which were impacted upon *Kat2a* loss, where the analysis identified an alteration specifically in mitochondrial bioenergetics and ribosomal biosynthetic programmes.

Chapter-5 discusses the mechanistic contribution of transcriptional programmes altered during pre-leukaemia transformation in *RUNX1-RUNX1T1(9a), Idh1*R132H and contrasts it with *MLL-AF9*. Specifically, the chapter describes the analysis of self-renewal potential and clonogenic capacity in *MLL-AF9* transformed primary cells segregated on the basis of high and low mitochondrial mass. The analysis suggested that *Kat2a* WT cells with low mitochondrial content may phenocopy some of the characteristics of *Kat2a* NULL cells. This chapter also details the analysis of protein synthesis activity in pre-leukaemia models using the OP-Puro incorporation method. A reduction in protein synthesis was observed in *Idh1*R132H *Kat2a* NULL cells. Further, upon inhibition of protein synthesis using S6K1 inhibitor, the *Kat2a* WT cells transformed with either *RUNX1-RUNX1T1(9a)* or *Idh1*R132H showed an enhanced colony forming potential, thus phenocopying *Kat2a* NULL cells. This contrasted with findings in the *MLL-AF9* model of leukaemia, where a reduction in colony forming potential was observed.

Chapter-6 discusses the role of gene expression variability consequent to *Kat2a* loss in deciding lineage dependencies and cellular fate reprogramming during pre-leukaemia progression. The single-cell RNA sequencing data described in Chapter-4 was used for this analysis. The analysis indicated that *Kat2a* NULL cells displayed an increase in transcriptional variability, which was accompanied by diversification of cell fates towards B-lymphocytes and monocytes. Furthermore, pseudo-temporal ordering of single *Kat2a* NULL cells revealed a highly branched trajectory, which was heavily populated with cells at intermediate stages of transformation; including accumulation of leukaemia progenitors with *RUNX1-RUNX1T1* signature. In contrast, *Kat2a* WT cells displayed a linear normal haematopoiesis trajectory with minimal branching and an abrupt transition towards candidate leukaemia progenitor state. This chapter also discusses single-cell ATAC sequencing analysis of a representative RUNX1-RUNX1T1(9a) human cell line, which suggested that the observed increase in transcriptional variability was a consequence of differential chromatin accessibility patterns observed upon inhibition of KAT2A.

Finally, all these findings are discussed together in Chapter-7 where they are critically evaluated to draw the main conclusions about how transcriptional variability consequential to differential epigenetic landscape may facilitate cell-fate dynamics during the process of leukaemia initiation.

2 Materials and Methods

2.1 *Kat2a* conditional knock-out model

Kat2a$^{fl/fl}$ mouse strain was a kind gift from Prof. Sharon Dent, MD Anderson Cancer Centre, Smithville, TX, USA. My lab generated a stable mouse line homozygous for the *Flox* allele, by crossing *Kat2a*$^{fl/fl}$ conditional knockout mice (MGI:3801321) (Lin *et al.*, 2008) with interferon response-inducible *Mx1-Cre*$^{+/-}$ transgenic mice (Kühn *et al.*, 1995a) in a C57Bl/6 background (Fig 2.1). *Kat2a* locus excision was obtained through the treatment of experimental (*Kat2a* excised-NULL) and control (*Kat2a* floxed- *Kat2a* WT) mice with intra-peritoneal polyinosylic-polycytidylic (pIpC) acid (Chan *et al.*, 2011).

After generating *Kat2a*$^{fl/fl}$ conditional knockout mouse model, my lab further confirmed that *Kat2a* deletion does not perturb normal haematopoiesis and thus preserves candidate progenitor cells-of-origin for leukaemia transformation (Domingues *et al.*, 2020).

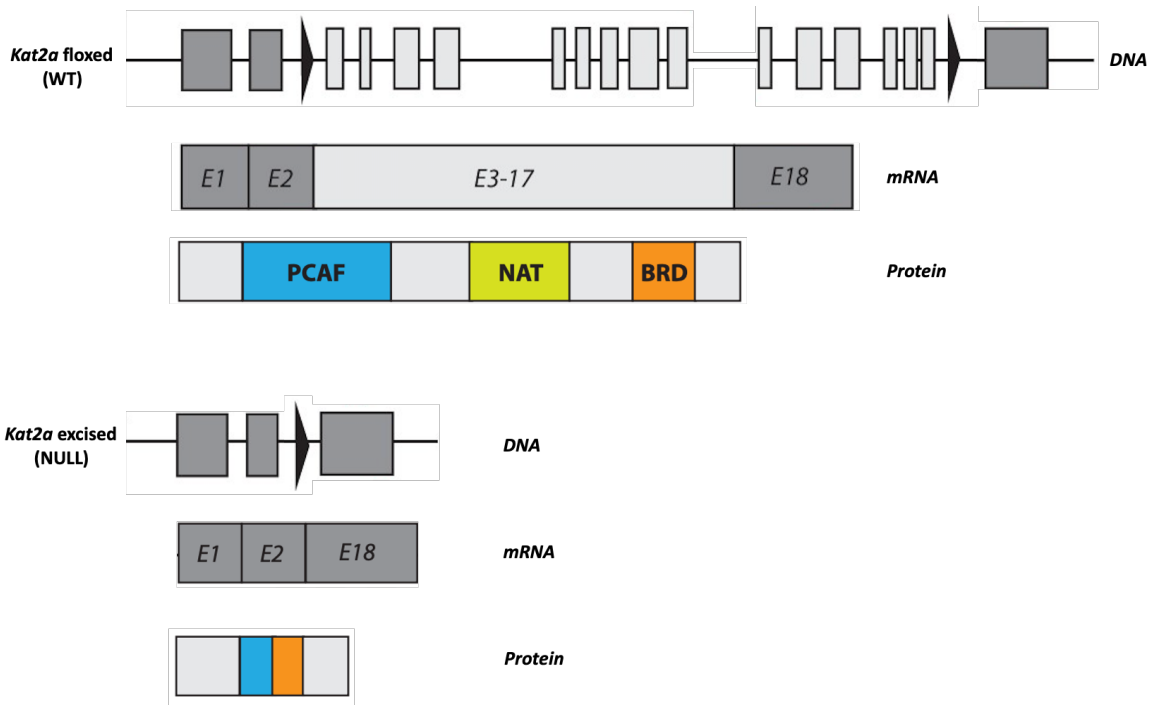


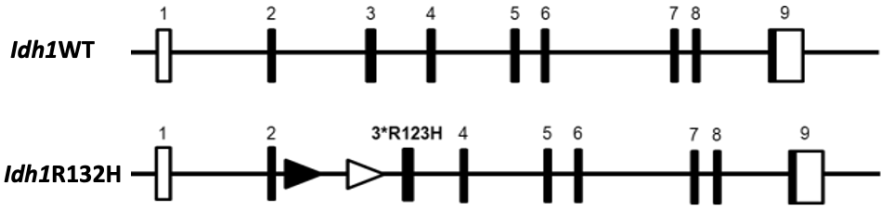
Figure 2.1*: Kat2a*$^{fl/fl}$ conditional knockout mouse model.

Schematic representing conditional *Kat2a* floxed (WT) and *Kat2a*-excised (NULL) alleles and resultant transcript and protein generated. PCAF- p300/CBP-associated factor, NAT- N-terminal and Ada-Two interaction domain, BRD- Bromodomain.

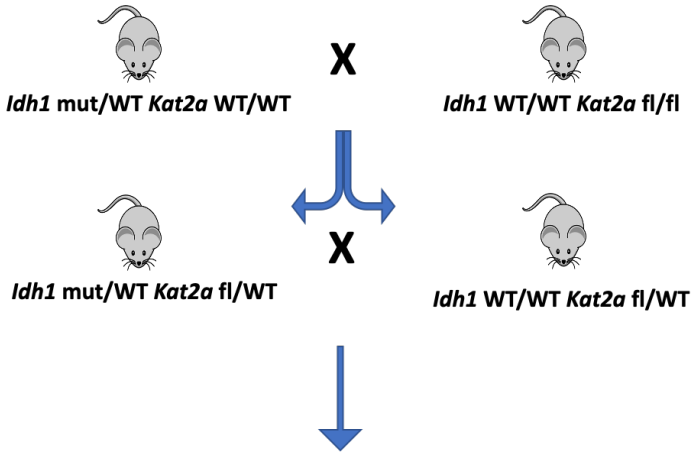
2.2 Generation of *Idh1*R132H *Kat2a* fl/fl mice model

Mx1-Cre inducible *Idh1*R132H mouse model (Fig 2.2A) was a kind gift from our collaborator Prof. George Vassiliou, Wellcome Sanger Institute, UK.

A



B



		Kat2a WT/WT	Kat2a fl/WT	Kat2a fl/fl
Idh1	WT/WT	Idh1 WT/WT Kat2a WT/WT	Idh1 WT/WT Kat2a fl/WT	Idh1 WT/WT Kat2a fl/fl
	mut/WT	Idh1 mut/WT Kat2a WT/WT	Idh1 mut/WT Kat2a fl/WT	**Idh1 mut/WT Kat2a fl/fl**

Kat2a HET- *Idh1* mut/WT *Kat2a* fl/WT

Kat2a NULL- *Idh1* mut/WT *Kat2a* fl/fl

Figure 2.2: Generation of *Idh1*R132H *Kat2a* fl/fl model.
(A) *Idh1*R132H conditional knock-in mouse model representing *Mx1-Cre* inducible R132H mutation in exon 3, **(B)** Schematic representing crossing strategy between *Idh1* mut/WT *Kat2a* WT/WT and *Idh1* WT/WT *Kat2a* fl/fl leading in order to obtain *Kat2a* Het- *Idh1* mut/WT *Kat2a* fl/WT and *Kat2a* NULL- *Idh1* mut/WT *Kat2a* fl/fl.

In order to generate *Mx1-Cre* inducible mouse colony with *Idh1*mut/WT and *Kat2a* floxed allele, *Idh1*mut/WT *Kat2a* WT/WT Cre +/- male was crossed with *Idh1* WT/WT *Kat2a* fl/fl Cre -/- female. The first-generation offspring obtained had a *Kat2a* fl/WT genotype (referred as *Kat2a* Het) with either *Idh1* WT/WT or *Idh1*mut/WT and a heterozygous Cre allele.

These off-springs having *Idh1* WT/WT *Kat2a* fl/WT and *Idh1*mut/WT *Kat2a* fl/fl were crossed in order to obtain experimental genotypes referred to as *Kat2a* HET- *Idh1*mut/WT *Kat2a* fl/WT and *Kat2a* NULL- *Idh1*mut/WT *Kat2a* fl/fl with heterozygous Cre allele (Fig 2.2B).

2.3 Genotyping

2.3.1 DNA Extraction

Ear clippings were collected from 4-5 weeks old mice and immediately stored in 500µl of lysis buffer having following composition-

50mM Sodium Chloride (NaCl)
50mM Tris(hydroxymethyl)aminomethane hydrochloride (Tris-Cl)
5mM Ethylenediaminetetraacetic acid (EDTA)
20% Sodium dodecyl sulphate (SDS)
0.5mg/ml Proteinase K

The lysis step was performed overnight at 55°C, 750rpm using a Thermoshaker (BioSan). Following day, the lysates were centrifuged at 13,000rpm for 5 minutes (tabletop centrifuge, eppendorf). The supernatant obtained was transferred to a fresh 1.7ml microtube having 500µl of Isopropanol. The microtubes were shaken gently until a white precipitate was observed and centrifuged at 13,000rpm for 5 minutes. The supernatant was discarded and the pellet was

washed with 700µl of 70% ethanol. The pellet was centrifuged twice at 13,000rpm for 5 minutes to remove excess ethanol and was further air-dried for 20 minutes and then resuspended in 20µl of Tris-EDTA (TE) buffer (10mM Tris, 1mM EDTA). The resuspended pellet was then incubated at 55°C, 750rpm for 10 minutes. The DNA obtained was quantified using Nanodrop™ Lite Spectrophotometer (Thermo Scientific) and stored at -20°C for long term storage.

2.3.2 Polymerase Chain Reaction for Mx1-Cre, Idh1 and Kat2a

The DNA samples extracted from the above step was utilised for Polymerase Chain Reaction (PCR) analysis using primers mentioned in Table B.2 (Annexure-B). The PCR master mix was prepared with the following composition-

7.5µl HotStarTaq Plus Master Mix 2X (Qiagen)
0.1µl Forward Primer 100µM
0.1µl Reverse Primer 100µM
2µl DNA template (50ng equivalent)
5.3µl RNase free water (Qiagen)

The PCR reaction set-up was as follows-

Initial Denaturation- 95°C for 5 minutes
Denaturation- 94°C for 30 seconds
Annealing- 60°C for Mx1-Cre and *Kat2a*, 57°C for Idh1 for 30 seconds
Elongation- 72°C for 30 seconds for *Kat2a* and Idh1, for 90 seconds for Mx1-Cre
} 40 cycles
Final Elongation- 72°C for 10 minutes

After PCR reaction was completed, 15µl of PCR product was mixed with 4X gel loading dye and loaded on a 1.5% agarose gel (Fig 2.3A, B). The band sizes were estimated based on 1kb Plus DNA ladder (New England BioLabs). The following were the corresponding band size for each genotype-

Mx1-Cre -/- ~1kb

Mx1-Cre +/- ~700bp and ~1kb

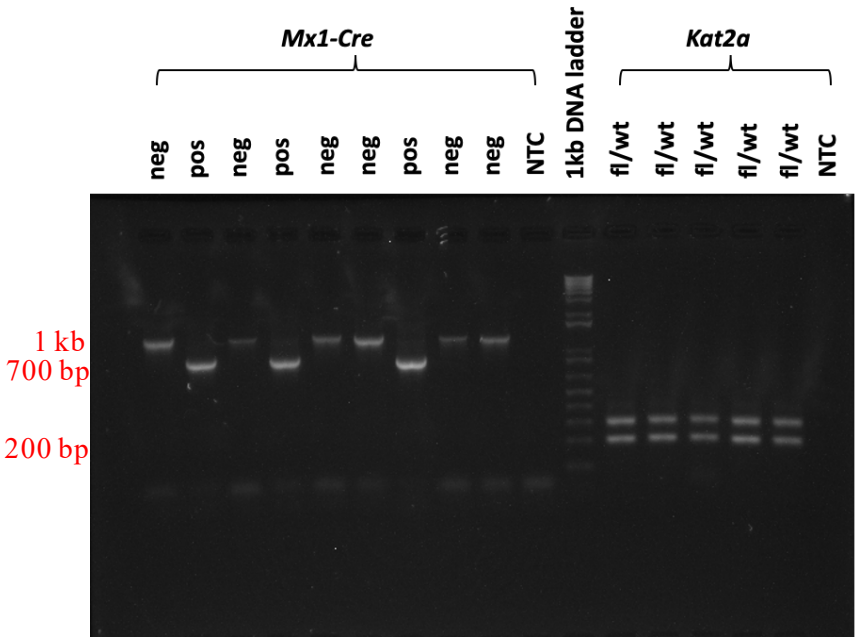
Kat2a fl/fl- 310bp

Kat2a fl/wt- 310bp and 210bp

Idh1 wt/wt - 675bp

Idh1 mut/wt- 750bp and 400bp

A



B

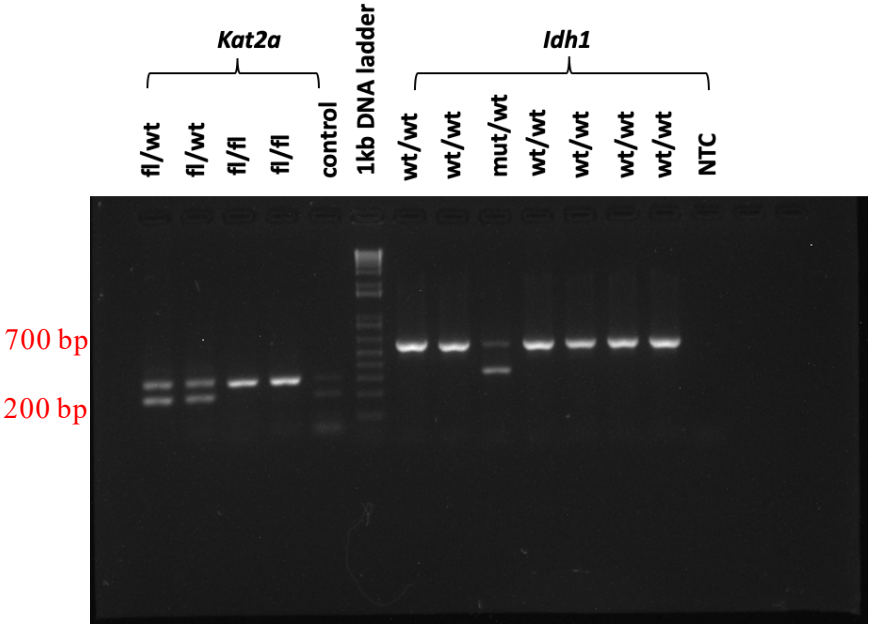


Figure 2.3: Representative gel images for genotyping from a single run.
(A) Gel image representing PCR products for 9 mice for Mx1-Cre and *Kat2a* where NTC represents No Template Control, **(B)** Gel image representing PCR products for *Kat2a* and *Idh1*.

2.4 *Idh1*R132H recombination confirmation

In order to confirm the recombination event leading to the activation of *Idh1*R132H, peripheral blood sampling was conducted post *pIpC* injections, using saphenous vein method (Fig 2.4A). Blood sampling was done every 2 weeks post *pIpC* injections and DNA extraction was performed as per manufacturer's instructions (PureLink Genomic DNA Mini Kit, Thermo Fisher Scientific). DNA was eluted in 20 µl of elution buffer and quantified using Nanodrop™ Lite Spectrophotometer (Thermo Scientific).

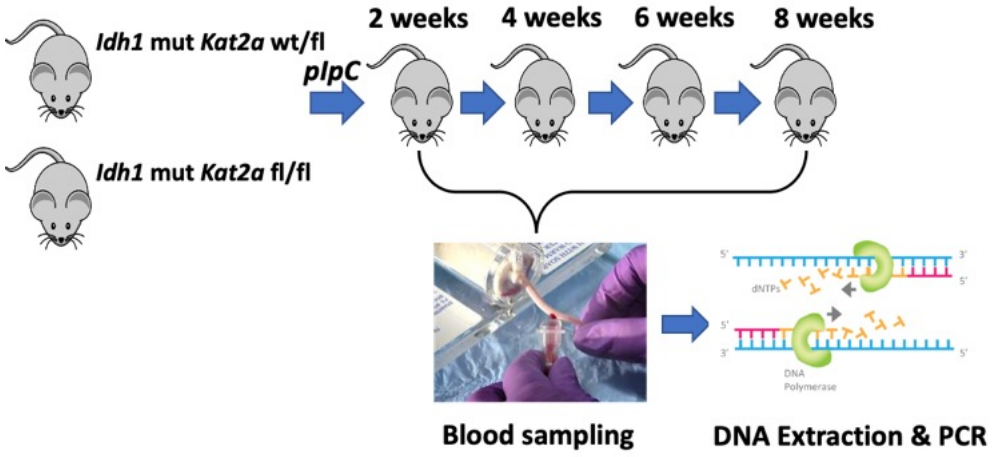
Next to assess the recombination event, 25ng of DNA sample was taken for PCR reaction (total volume- 15µl) along with primers designed specifically for *Idh1*R132H mutation (Table B.3 (Annexure-B)).

7.5µl HotStarTaq Plus Master Mix 2X (Qiagen)
0.1µl Forward Primer 100µM
0.1µl Reverse Primer-1 100µM
0.1µl Reverse Primer-2 100µM
2µl DNA template (50ng equivalent)
5.2µl DNase RNase free water (Qiagen)

The PCR reaction set-up was as follows-

Initial Denaturation- 95°C for 5 minutes
Denaturation- 94°C for 30 seconds
Annealing- 57°C for 30 seconds ┐ 40 cycles
Elongation- 72°C for 30 seconds ┘
Final Elongation- 72°C for 10 minutes

A


Idh1 mut Kat2a wt/fl
Idh1 mut Kat2a fl/fl
pIpC
2 weeks
4 weeks
6 weeks
8 weeks
Blood sampling
DNA Extraction & PCR

B

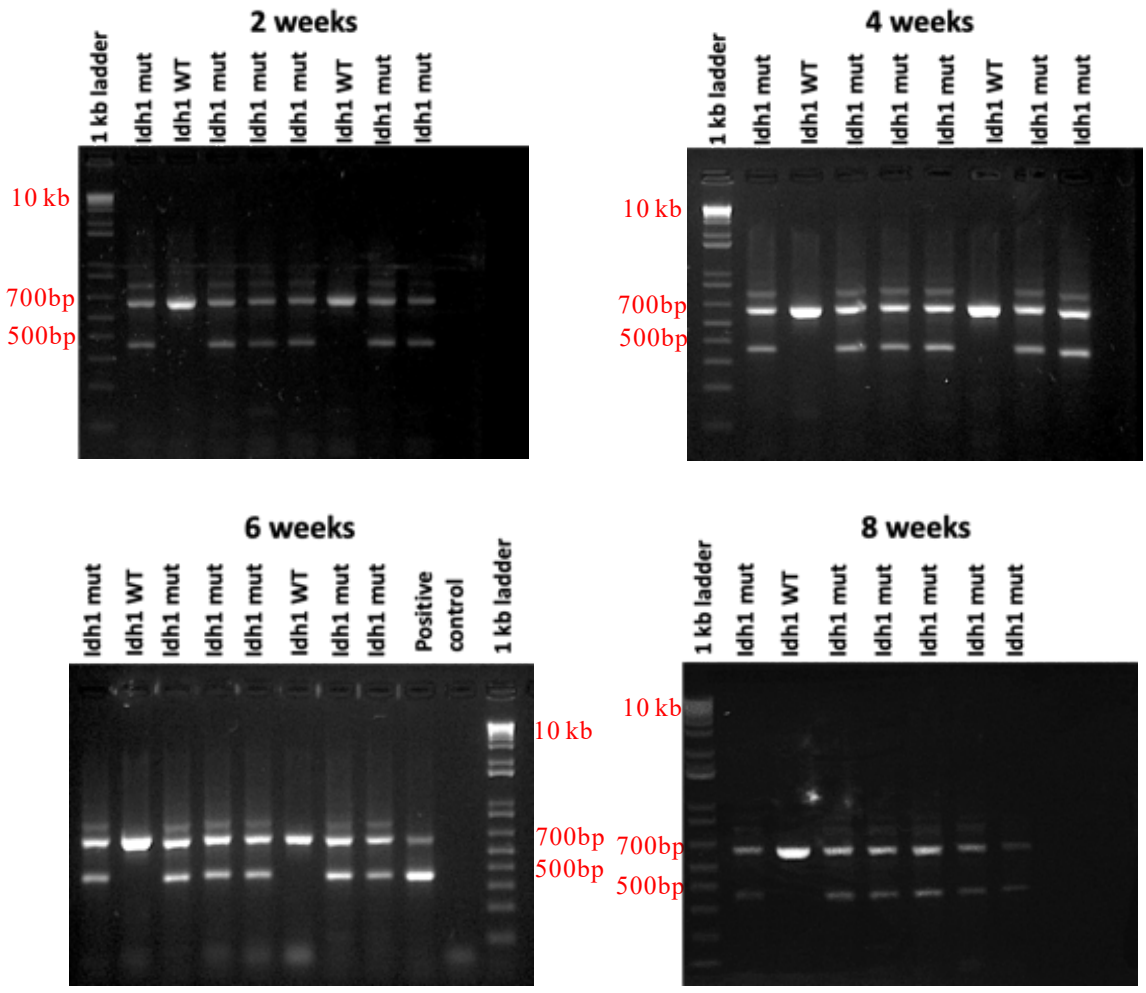

2 weeks
1 kb ladder
Idh1 mut
Idh1 WT
10 kb
700bp
500bp
4 weeks
6 weeks
Positive control
8 weeks

Figure 2.4: *Idh1*R132H recombination analysis.
(A) Schematic for peripheral blood sampling post 2/4/6 and 8 weeks of *pIpC* treatment. DNA extraction was performed from peripheral blood and PCR was performed to confirm the activation of *Idh1*R132H mutation, (**B)** Representative gel electrophoresis images for *Idh1*R132H at different sampling time points.

After PCR reaction was completed, 15µl of PCR product was added to 4X gel loading dye and loaded on a 1.5% agarose gel. The band sizes were visualized using AlphaImager UV transilluminator (Protein Simple) estimated based on 1kb Plus DNA ladder (New England BioLabs).

A recombination event highlighting activation of *Idh1*R132H mutation corresponds to an additional band size of ~700bp (Fig 2.4B).

2.5 *Kat2a* excision confirmation

To confirm the excision levels of *Kat2a* floxed allele, the DNA samples obtained from the above step were utilised for Quantitative Reverse Transcription Polymerase Chain Reaction (qRT-PCR). The following reaction was prepared in triplicates for each sample-

10µl 2X Takyon Low Rox SYBR Mastermix dTTP Blue (Eurogentec)
0.1µl Forward Primer 100µM
0.1µl Reverse Primer 100µM
2µl DNA template (40 ng equivalent)
7.8µl RNase- free water

Primers were designed in two different pairs, where one pair was designed to amplify the excised region of *Kat2a*, referred to as *Kat2a*-IN, whereas the second pair was designed to amplify the region distal to the second loxP site, referred to as *Kat2a*-OUT (Fig 2.5A) (Table B.3 (Annexure-B)). Standards were prepared ranging from 0.8ng- 80ng along with a no template control. The reaction conditions were as follows-

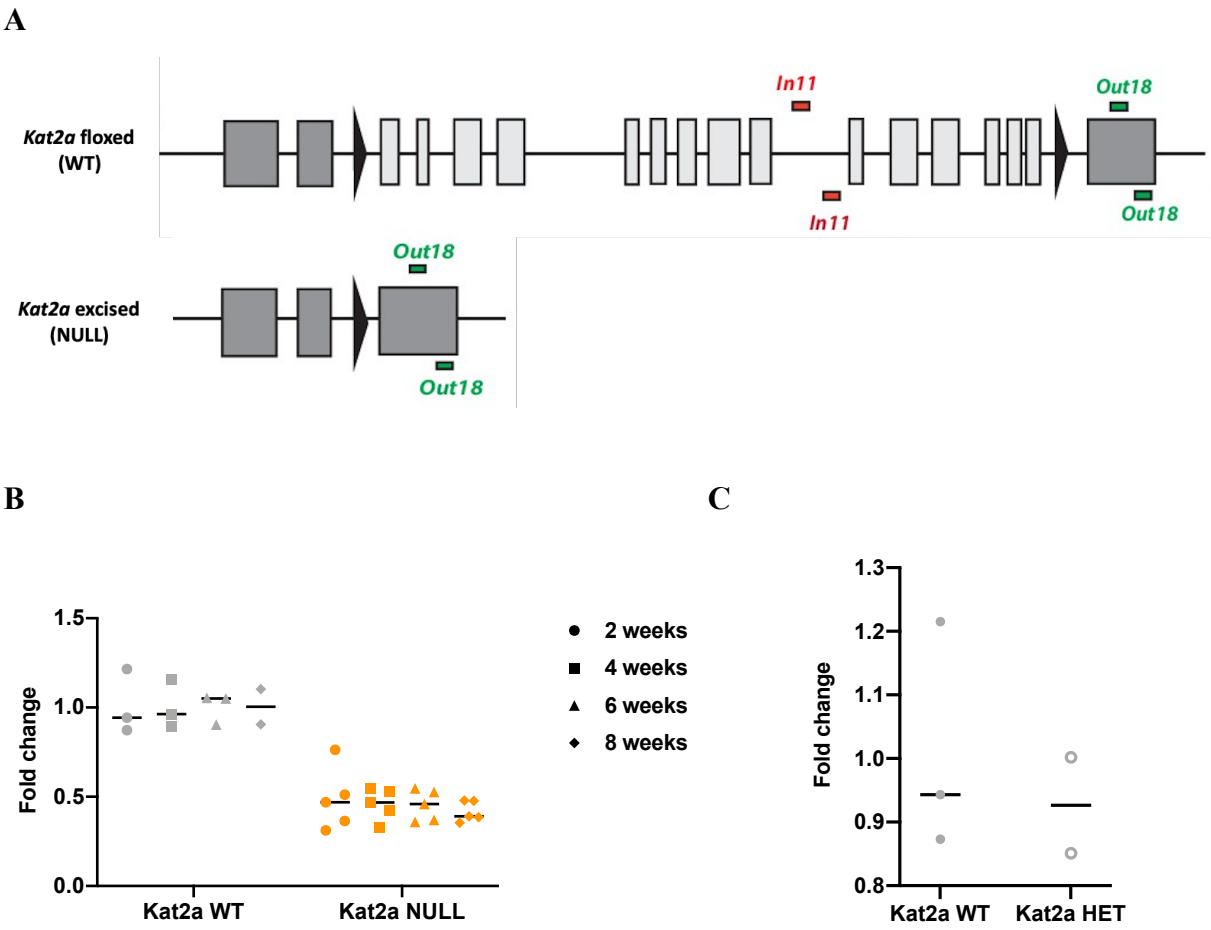


Figure 2.5: *Kat2a* excision analysis.

(A) Primer designing strategy depicting WT and *Kat2a* fl/fl to confirm *Kat2a* excision *Kat2a* IN (In11, within the excised region) and *Kat2a* OUT primers (Out18, downstream of the excised region), **(B)** qRT-PCR analysis to confirm *Kat2a* excision from peripheral blood sample (n=3 for *Kat2a* WT and n=5 for *Kat2a* NULL), **(C)** qRT-PCR analysis comparing excision between *Kat2a* WT (n=3) with *Kat2a* Het (n=2).

Initial denaturation- 95°C for 3 minutes

Denaturation- 95°C for 10 seconds } 40 cycles

Annealing- 60°C for 1 minute

Final cycle- Denaturation 95°C for 60 seconds, Annealing 55°C for 30 seconds

The expression levels were calculated by the Pfaffl method (Pfaffl, 2001) following normalization with respect to *Kat2a*-OUT (Fig 2.5B, C).

2.6 Bones and spleen processing

Experimental animals were culled by cervical dislocation method, one of the schedule 1 method of euthanasia. Before dissection, the animal was cleaned by spraying 70% ethanol. Animal was dissected, flesh was removed using forceps and scalpel in order to collect leg bones, femur, tibia and spleen in I10 (Iscove's Modified Dulbecco's Medium (IMDM) supplemented with 10% Heat inactivated- Fetal Bovine Serum (FBS), 2mg/mL L-Glutamine, 1% Penicillin-Streptomycin Amphothericin (PSA)) for further processing. For histological assessment, liver, lungs, sternum, a portion of spleen, heart, kidneys, intestine were collected, weighed and immediately stored in formalin.

Extraction of bone marrow (BM) and spleen cells was done separately using mortar and pestle (Fig 2.6). The resultant extract was filtered using 40µM cell strainer (Fisher Scientific) to obtain a homogeneous suspension. The cells were centrifuged at 2000rpm for 5 minutes (refrigerated centrifuge, eppendorf). Post removal of the supernatant, cells were resuspended in 2ml of Red Blood Cell (RBC) lysis buffer and incubated for 5 minutes. The reaction was terminated by adding 10ml of I10 medium. The cells were again centrifuged at 2000rpm for 5 minutes. The supernatant was decanted and cells were resuspended in 20ml of Phosphate Buffer EDTA containing PBS + 2mM Ethylenediaminetetraacetic Acid (EDTA) + 0.5% Bovine Serum Albumin (BSA). The cells were counted, of which $1*10^6$ cells/ sample were kept for flow cytometry analysis (described below), 50,000 cells/ sample were utilised for colony-forming assay (described below). Rest of the cells were cryopreserved with two vials each for BM and spleen. For cryopreservation, cells were centrifuged at 2000rpm for 5 min, supernatant was removed and the pellet was resuspended in 1ml of I10 medium with 1ml of freezing medium having 80% FBS and 20% Dimethylsulfoxide (DMSO). Each cryovial containing 1ml of this cell suspension was stored in Mr. Frosty freezing container (Thermo Scientific) at -80°C for up to 3 days and transferred to liquid nitrogen tank for long term storage.

For pre-leukaemia studies, animals were processed in the same manner at pre-decided time points post transplantation. Leukaemia was determined on the basis of the presence of clinical

symptoms of hunched posture, inappetence and lethargy, and animals were processed as mentioned above.

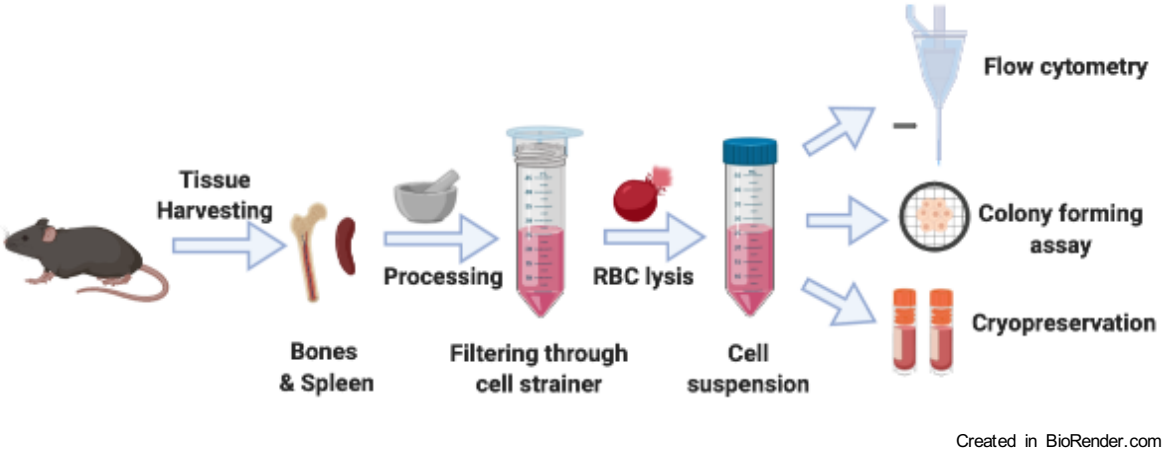


Figure 2.6: Tissue Processing.

Schematic representing the processing of tissue samples obtained post Schedule 1 and further used for *in vitro* experiments.

2.7 Lineage depletion

Lineage depletion was done to enrich for progenitor population of cells using a cocktail of biotinylated antibodies including B220, Ter119, Cd11b, Gr1 and Cd3e as mentioned in Table B.4 (Annexure-B). For this, primary BM cells obtained from mouse were resuspended in 100µl of PBE and a total of 10µl lineage cocktail antibodies were added to the cell suspension. The cell suspension having antibodies was incubated for 15 minutes on ice. Post incubation, the cell suspension was centrifuged at 2000rpm for 5 minutes and supernatant was removed. The pellet was resuspended in 100µl PBE along with magnetic Streptavidin Nanobeads (BioLegend) at a ratio of 10µl of beads for every 10^7 cells. The cell suspension was incubated with beads for 15 minutes on ice and washed post incubation. The pellet was resuspended in 5ml of ice-cold PBE and cell suspension was transferred to polystyrene tube (Fisher Scientific) compatible with magnetic separation. The tube was placed inside the magnet (MojoSort Magnet, BioLegend) for 5 minutes. Post incubation, the cell suspension was poured out and collected. The magnetic separation steps were repeated in order to obtain a pool of two separations. Cells obtained were counted and further utilised for retroviral transduction.

2.8 Retroviral Transduction

For overexpression of *RUNX1-RUNX1T1(9a)* fusion protein, retroviral particles expressing the fusion protein were produced using Human Embryonic Kidney (HEK) 293T cells. For this, $2.5*10^6$ of HEK 293T cells were cultured in a 10 cm dish in D10 (DMEM supplemented with 10% Hi- FBS, 2mg/mL L-Glutamine, 1% PSA) at 37°C in CO_2 incubator for overnight. Next day, the cells were transfected using the following transfection mix for each plate-

47.5µl of TransIT (Mirus)
5µg of psi Eco vector (packaging plasmid)
5µg of *RUNX1-RUNX1T1(9a)* retroviral plasmid (*MSCV-AML1/ETO-IRES-GFP*)
600µL of Opti-MEM Medium (Gibco)

This transfection mix was added dropwise to HEK 293T cells followed by plate swirling and cultured overnight at 37°C in CO_2 incubator. Following the day of transfection, the medium was replaced with R20 (Rosewell Park Memorial Institute 1640 (RPMI-1640) medium supplemented with 20% Hi- FBS, 2mg/mL L-Glutamine, 1% PSA, Table B.5 (Annexure-B)). After 24 hours and 48 hours of medium replacement, medium was collected and filtered using 0.45µM syringe filter (Merck Millipore). The viral supernatant collected was immediately used for transduction or stored at -20°C for long term storage.

For viral transduction, BM cells freshly isolated/ thawed were maintained at a density of $1*10^6$ cells/ml in R20 supplemented with 20ng/ml of mouse-Stem Cell Factor (mSCF), 10ng/ml of mouse Interleukin-3 (mIL-3), 10ng/ml of mouse Interleukin-6 (mIL-6) (Table B.5 (Annexure-B)) in a 6-well plate for overnight at 37°C, 5% CO_2. The next day, BM cells were briefly centrifuged at 2000 rpm for 5 minutes and viral supernatant supplemented with 20ng/ml of mSCF, 10ng/ml of mIL-3, 10ng/ml of mIL-6 and 10µg/ml of polybrene (Sigma) (Table B.5 (Annexure-B)) was added to cells at a density of $1*10^6$ cells/ml. Post viral addition, cells were centrifuged at 2000 rpm at 25°C for 1 hour and kept overnight at 37°C 5% CO_2. A second round of viral transduction was performed the following day in a similar manner. Post two rounds of viral transductions, cells were collected, washed twice with PBS and once with R20 medium.

The cell pellet obtained was resuspended back in R20 medium and GFP levels were assessed by Gallios Flow Cytometer (Beckman Coulter) and data was analysed using Kaluza software.

2.9 *RUNX1-RUNX1T1(9a)* experiment set-up

For *RUNX1-RUNX1T1(9a)* leukaemia and pre-leukaemia studies, three animals each of *Kat2a* WT and *Kat2a* NULL cohort were intraperitoneally injected with *pIpC* (Sigma) for 5 alternate days (10 days duration) at 300 µg/injection. After 4-6 weeks of the last *pIpC* injection, animals were culled following schedule 1 method of euthanasia and BM cells were isolated from individual animals as mentioned above. The BM cells obtained were enriched for lineage depletion markers and retrovirally transduced with *RUNX1-RUNX1T1(9a)* overexpressing plasmid as described above. Post transduction, Green Fluorescent Protein (GFP) levels, representative of transduction efficiency, were assessed using Gallios Flow Cytometer (Beckman Coulter).

For the experimental set-up, separate pool of *Kat2a* WT and *Kat2a* NULL BM cells were made post *in vitro* transduction from three animals each. C57/BL6 mice, >8 weeks old, were lethally irradiated (2*5.5 Gy), weighed and injected with $1*10^6$ cells/recipient from WT or NULL pool cells (n= 17 mice/group) intravenously. Pre-leukaemic mice were processed 2 months (n= 3/group) and 4 months (n= 2/group) post-transplantation. Leukaemia studies were performed on the rest of the transplants where mice were collected based on the presence of clinical symptoms of hunched posture, inappetence and lethargy.

Mice were kept in a Specific Pathogen Free (SPF) animal facility, and all experimental work was carried out under UK Home Office regulations. Animal research was regulated under the Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012 following ethical review by the University of Cambridge Animal Welfare and Ethical Review Body (AWERB).

2.10 Flow Cytometry analysis

Cell surface staining for both BM and spleen obtained from pre-leukaemia and leukaemia animals was performed using a panel of antibodies which included Cd117/ c-Kit-APCC7,

Cd11b/ Mac1-AF700, Sca1-PEC7, Gr1-PB, Cd34-APC and Cd16/32/Fcγr-PE. The dilutions used and respective clones for each antibody is mentioned in Table B.4 (Annexure-B). For this, $1*10^6$ cells/ sample were resuspended in 100µl of PBE and antibody cocktail was added. Unstained sample and respective single colour controls were prepared to define positive gates. The samples were incubated for 20 minutes on ice and the reaction was terminated by adding 500µl of R20 media. The samples were then centrifuged at 2000 rpm for 5 minutes and further resuspended in 200µl of R20 media for analysis. Acquisition was performed on Gallios Flow Cytometer (Beckman Coulter), gating was done on GFP^+ population and data was analysed using Kaluza software.

For surface marker studies during *in vitro* transformation of *Kat2a* WT and NULL with *RUNX1-RUNX1T1(9a)*, same protocol was followed for a panel of antibodies including B220-APCCy7, F4/80 PE, CD14-PECy7 (BioLegend). The cells were acquired on Attune NxT Flow Cytometer (Thermo Scientific) and analysed using Attune Nxt Software version 3.2.1.

2.11 Colony Formation assay

For analysis of pre-leukaemia samples from *RUNX1-RUNX1T1(9a)* and *Idh1*R132H, 50,000 cells/ condition isolated from BM were plated in duplicates in MethoCult M3434 (Stem Cell Technologies) and colonies (a colony is defined as the presence of more than 25 cells present in close proximity to each other) were scored after 7-10 days. The colonies were categorized as Granulocyte (G), Erythroid (E), Macrophage (M), Granulocyte Macrophage (GM) and Granulocyte Erythroid Macrophage Megakaryocyte (GEMM). Further, re-platings were done with 10,000 cells/condition in duplicates until no colonies were discovered in order to assess the self-renewal potential of transformed cells.

In case of *MLL-AF9* transformed primary mouse BM cells, 10,000 cells/ condition were plated as mentioned above and colonies were scored after 5-7 days. The colonies were categorized as compact, mixed and dispersed. Further re-platings were done with 2,000- 4,000 cells/ condition to obtain transformed cells.

2.12 Peripheral Blood analysis

To study the progression of leukaemia, blood sampling was done every 4-5 weeks post transplantation using saphenous vein method. To start with, animals were warmed at 37°C in a temperature-regulated chamber for 10-15 minutes in order to dilate the blood vessel before bleeding. After this, the individual animal was placed in a restraining device in order to access the tail vein. The vein was punctured using 30G needle (VWR) and sample was collected in capillary action collection tube (Sarstedt). The samples collected were analysed for different haematological parameters using a Vet abc automated counter (Scil Animal Care, Viernheim, Germany).

For analysing GFP population, 50µl of blood sample was diluted in 1:1 ratio with PBS. RBC lysis was performed for 10 minutes and cells were resuspended in 100µl of PBS post centrifugation at 200rpm for 5 minutes. The cells were acquired on Gallios Flow Cytometer (Beckman Coulter) and data was analysed using Kaluza software.

2.13 Maintenance experiment

To study the self-renewal potential of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT and *Kat2a* NULL cells in a maintenance set-up, bone marrow cells were collected from 12 weeks old *Kat2a* -/- *Mx1-Cre* -/- animals and the cells obtained from three such animals were pooled together. The pooled cells obtained were then retrovirally transduced with *RUNX1-RUNX1T1(9a)* as mentioned above. Post retroviral transduction, three rounds of plating were done as colony forming assay with 4000 cells utilised for each plating. Re-platings were done 10 days post previous plating. Post plating 2, the cells were collected and then transduced retrovirally with MIGR-Cre-OP-Puro where MIGR-OP-Puro served as control, as mentioned above. The transduced cells were then selected using puromycin selection for 48 hours in liquid culture medium. The selected cells were then again plated as colony forming assay with 4000 cells/condition in the presence of puromycin. The number of colonies obtained were scored and reported.

2.14 *Idh1*R132H experiment set-up

For *Idh1R132H* pre-leukaemia and leukaemia studies, 8-weeks old mice with required genotypes i.e., *Idh1R132H Kat2a* wt/fl (*Kat2a* HET) and *Idh1R132H Kat2a* fl/fl (*Kat2a* NULL) were subjected to *pIpC* injections intraperitoneally for 5 alternate days at 300 µg/dose. To study the phenotypic characterization of *Kat2a* HET and NULL carrying *Idh1*R132H during pre-leukaemia, mice were culled post 4-weeks and 20-weeks of *pIpC* injections following schedule 1 method of euthanasia (n= 3/ group for each time point) and BM cells were extracted using same procedure mentioned above. 50,000 cells/ condition were maintained in CFC assay with scoring and re-plating the colonies every 7-10 days as mentioned above. Rest of the samples were cryopreserved as described above.

For leukaemia studies, the BM cells collected post 20-weeks of *pIpC* injections were thawed and counted. $1*10^6$ cells/ recipient were injected into sub-lethally irradiated (1*5.5 Gy) CD45.1 animals (n= 8/group). Leukaemia was determined on the basis of the presence of clinical symptoms of hunched posture, inappetence and lethargy, and animals were processed as mentioned above.

2.15 *Idh1*R132H haematopoietic compartment staining

To study the impact of *Kat2a* NULL on different haematopoietic compartments in *Idh1*R132H mice, BM cells obtained from 4-weeks and 20-weeks post *pIpC* were thawed and recovered in R20 medium. $1*10^6$ cells/ sample was taken for cell surface staining whereas $0.2*10^6$ cells were taken for each unstained/ single colour control and Fluorescence minus one (FMO) for Cd135 and Cd34. Surface staining was performed using Cd16/32- FITC, Cd135-PE, Cd117/C-Kit-APCCy7, Sca1-PECy7, Cd34-APC as mentioned in Table B.4 (Annexure-B). Lineage exclusion cocktail included B220, Ter119, Cd11b, Gr1 and Cd3e biotinylated antibodies along with streptavidin-conjugated Brilliant Violet 510 as mentioned in Table B.4 (Annexure-B). The samples were acquired on Gallios Flow Cytometer (Beckman Coulter) and data was analysed using Kaluza software based on the gating strategy mentioned in Fig 2.7.

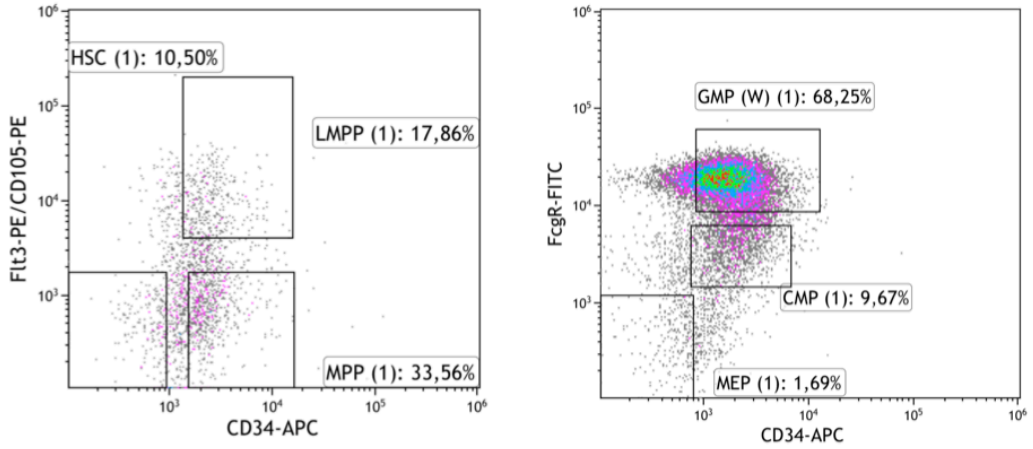


Figure 2.7: Flow cytometry analysis for normal haematopoietic compartments.
Gating strategy for different haematopoietic compartments in *Kat2a* HET and *Kat2a* NULL with *Idh1*R132H. The populations including HSC, LMPP, MPP, GMP, CMP and MEP are defined in Table B.1 (Annexure-B).

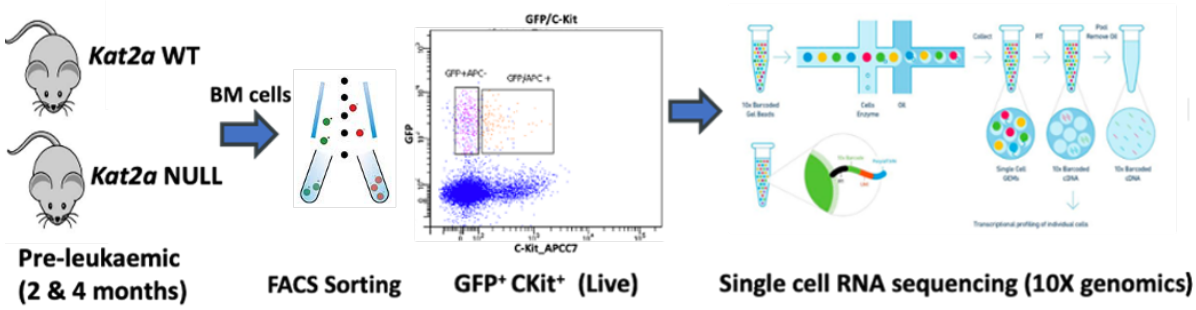
2.16 Single-cell RNA sequencing

2.16.1 Strategy and sample preparation

Single cell RNA sequencing (scRNA-seq) was performed for *RUNX1-RUNX1T1(9a)* pre-leukaemia BM samples collected post 2 months and 4 months of transplantation. For this, one representative BM sample of each genotype for individual time point was taken for further analysis. The samples were selected based on the percentage GFP population as well as self-renewing capacity as studied *in vitro* using CFC assay (Fig 2.8A).

These BM cells which were cryopreserved immediately after processing were thawed and recovered in R20 medium. In order to enrich for progenitor-like population, 10^7 cells/ sample were resuspended in 200 µl PBE and stained with Cd117/ c-Kit- APCCy7 for 30 minutes on ice. After incubation, cells were washed and resuspended in R20 medium with Hoescht58 (as mentioned in Table B.4 (Annexure-B)) in order to exclude dead cells. The stained cells were subjected to Fluorescence Activated Cell Sorting (FACS) for GFP^+Cd117^+ population based on the gating strategy in Fig 2.8B. The sorting was performed at NIHR Cambridge BRC Cell Phenotyping Hub situated in the Department of Medicine, Addenbrookes Hospital.

A


Kat2a WT
Kat2a NULL
BM cells
GFP/C-Kit
GFP+APC-
GFP/APC +
C-Kit_APCC7
Pre-leukaemic
(2 & 4 months)
FACS Sorting
GFP⁺ CKit⁺ (Live)
Single cell RNA sequencing (10X genomics)

B

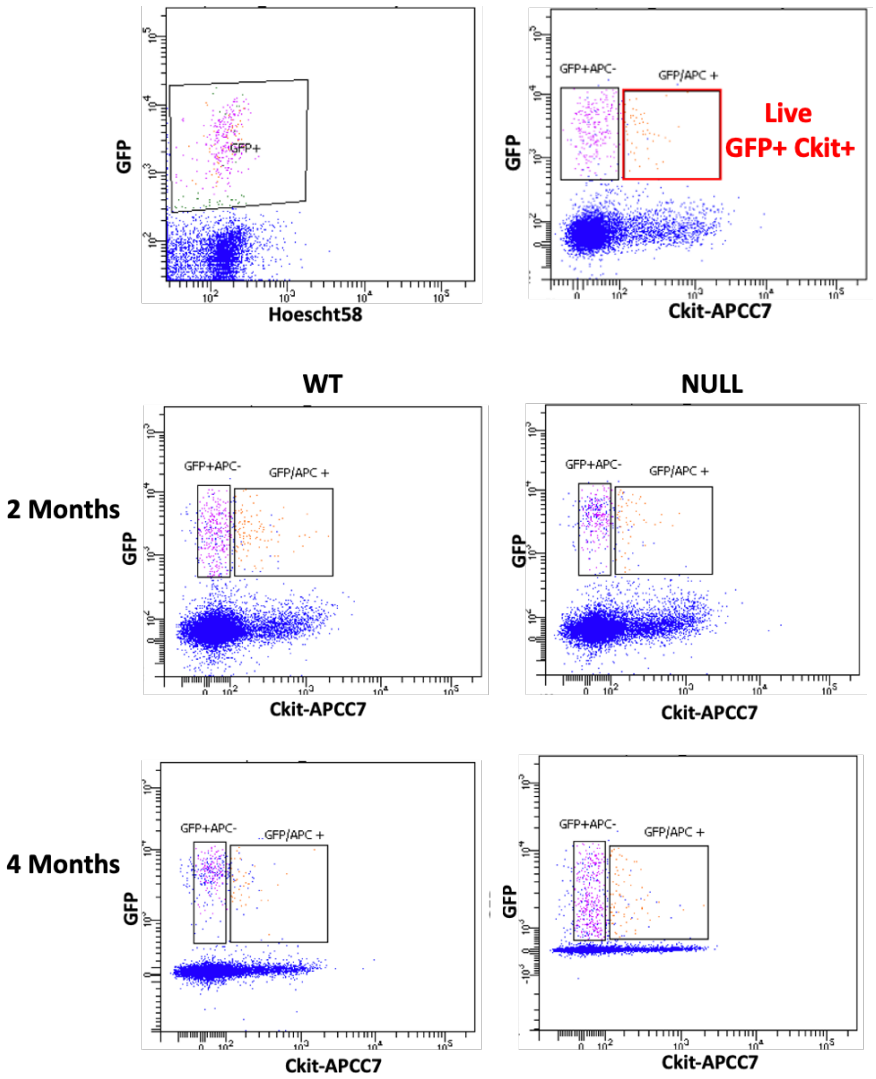

GFP
GFP+
Hoescht58
GFP+APC-
GFP/APC +
Live
GFP+ Ckit+
Ckit-APCC7
WT
NULL
2 Months
4 Months

Figure 2.8: Single cell RNA sequencing strategy and sample preparation.
(A) Sample preparation strategy for single-cell RNA sequencing obtained from *RUNX1-RUNX1T1(9a)* pre-leukaemia post 2 and 4 months of transplantation and sorted for GFP^+c-Kit^+ **(B)** Gating strategy for sorting mouse bone marrow cells for single cell RNA sequencing highlighted in red gate along with sorting layout for individual samples.

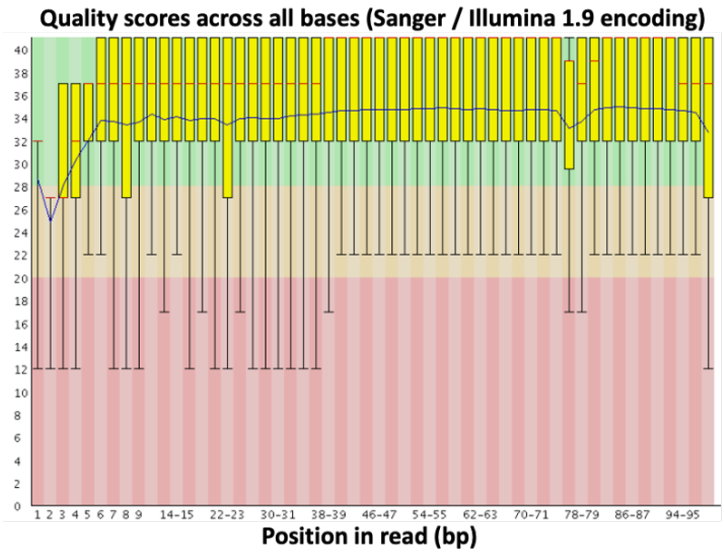
Each sample was sorted following the above gating strategy in order to obtain ~2000 cells/sample (Fig 2.8B). The sorted cells were collected in 500 µl of R20 medium. The cells were washed and counted post-sorting. 2000 cells/sample were resuspended in 34µl of R20 medium and sent for library preparation using 10X genomics technology at Genomics Core Facility in Cancer Research UK Cambridge Institute.

The library preparation involves the association of all cDNA molecules arising from a single cell with a unique barcode using gel bead-in emulsions. These barcodes are 10bp long and called unique molecule identifier (UMI). Broadly, the library preparation was done using 10X genomics 3' mRNA v2 chemistry with steps involving capturing cells with beads, release RNA, oligo- dT primed cDNA to get a 3' mRNA seq library. The libraries obtained were subjected to paired-end sequencing with 26 base pair (bp) read length using Hi Seq 4000.

2.16.2 Quality Control (QC) and generation of gene-cell matrix

In order to understand the quality of reads generated post sequencing, FastQC report generated post Illumina sequencing was studied where quality scores across all bases were observed using the following plots-

A



B

Figure 2.9: Quality Control.

(A) Plot representing quality score per base where x-axis represents the position in base pair and y-axis shows quality scores, **(B)** Plot representing the distribution of quality score over all sequences.

The plot in (Fig 2.9A) represents an overview of the range of quality values across all bases at each position in the FastQ file. For each position a BoxWhisker type plot is drawn where the central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points, the blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the

better the base call. The per sequence quality score report allows us to see if a subset of the sequences have universally low-quality values (Fig 2.9B). Ideally, these should represent only a small percentage of the total sequences.

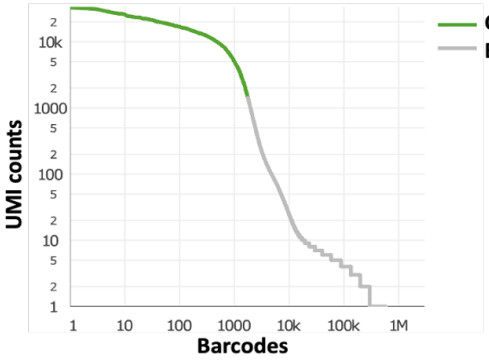
Post QC, Cell Ranger (v2.2) pipeline developed by 10X genomics was followed in order to align reads and generate gene-cell matrix. The process was initiated by cellranger mkfastq which was used to demultiplex raw base call (BCL) files generated by Illumina sequencing into FASTQ files. These FASTQ files obtained were further fed into cellranger count pipeline which performs alignment, filtering, barcode counting, and UMI counting. mm10 was used as a reference genome for alignment. These two pipelines were run individually for each sample in order to obtain good quality reads for further analysis.

For analysing all of the samples together, a combined gene- barcode matrix was generated using cellranger aggr pipeline (Fig 2.10A). This pipeline normalised all four samples to the same sequencing depth then recomputed the gene-barcode matrices and performed analysis on the combined data (Fig 2.10B, C, D). The gene- barcode matrix generated had 1767 cells in total 1575 median genes per cell and was used for further processing and analysis.

A

Parameter	WT-2M	NULL-2M	WT-4M	NULL-4M
Estimated no. of cells	379	369	518	501
Mean reads per cell	224927	235433	168460	205589
Median genes per cell	1916	1024	1925	1603
No. of reads	85247444	86875127	87262540	103000489
Valid barcodes	98.00%	97.80%	98.10%	98.10%
Reads mapped to genome	94.80%	94.00%	95%	94.50%
Reads confidently mapped to genome	90.50%	90.10%	90.40%	90.50%
Reads mapped confidently to intergenic regions	3.10%	4.50%	3%	3%
Reads mapped confidently to intronic regions	11.40%	15.90%	10.80%	10.50%
Reads mapped confidently to exonic regions	76.10%	69.70%	76.60%	76.90%
Reads mapped confidently to transcriptome	73.10%	67%	73.90%	74.00%
Reads mapped antisense to cell	1.20%	1.10%	1%	1.30%
Fraction Reads in Cells	88.10%	87.30%	90.90%	91.70%
Total Genes detected	13637	13379	13296	13258
Median UMI counts per cell	7488	2753	8270	6266

B



Estimated Number of Cells	1,767
Fraction Reads in Cells	89.8%
Pre-Normalization Mean Reads per Cell	205,085
Post-Normalization Mean Reads per Cell	174,770
Median Genes per Cell	1,575
Median UMI Counts per Cell	5,939

C **D**

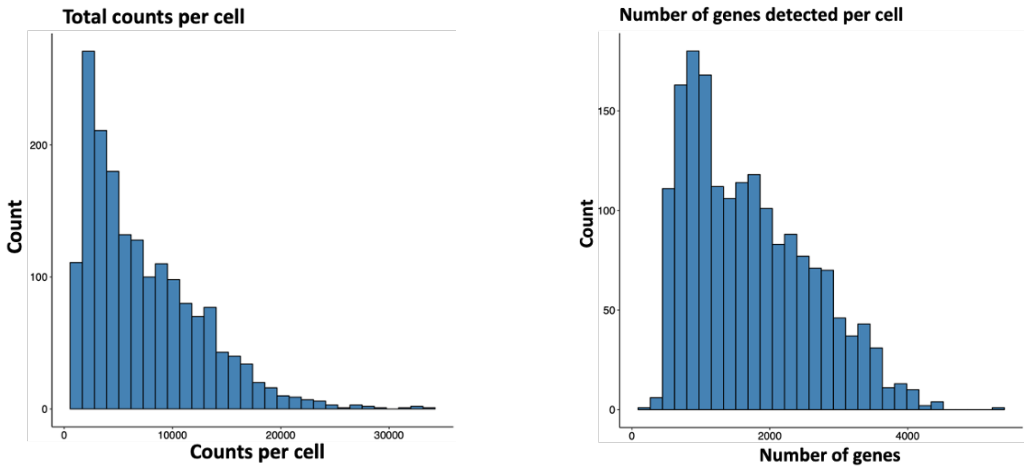


Figure 2.10: Generation of gene-cell matrix using Cellranger aggr pipeline.

(A) Tabular representation of different parameters corresponding to individual sample resulting from Cellranger aggr pipeline where median number of genes and UMIs per cell are highlighted in red, **(B)** Barcode versus UMI plot highlighting events considered as single cell in green whereas technical noise is considered background and highlighted in grey, **(C)** Distribution of total UMI counts per cell post normalization **(D)** Distribution of number of genes detected per cell post normalization.

2.16.3 Differential expression analysis

Differential expression analysis was done using DESeq2 tool (Love, Huber and Anders, 2014) integrated in Seurat v2.4. DESeq2 estimates variance-mean dependence and calculates differential expression based on a model using negative binomial distribution. For genes with adjusted p-value =< 0.05, differential expression was calculated using $\log_2$ fold change ($\log_2$ FC) where $|\log_2 FC| >= 0.26$ was considered a significant upregulation/ downregulation, corresponding to a 20% change in average fold difference.

2.16.4 Gene Ontology analysis

The differentially expressed genes obtained from DESeq2 analysis were further studied for biological processes enrichment. For this, Panther analysis (version 14.0) (H. Mi *et al.*, 2018) was done where enrichments were calculated based on Fischer's exact test and adjusted p-values were obtained upon Bonferroni correction for multiple testing.

2.16.5 Molecular Signature Database (MSigDB)

In order to understand the molecular function of the differentially expressed genes of respective WT and NULL comparisons obtained from DESeq2 analysis, an overlap of these genes with C2 (MF) datasets available on MSigDB v7.1 (Subramanian *et al.*, 2005, Liberzon *et al.*, 2015) was computed and top 10 gene sets with False Discovery Rate (FDR) less than 0.05 were represented.

For highlighting the processes specific for each cellular compartment, an overlap was performed with C7 dataset highlighting immunologic signatures and top 10 gene sets with FDR less than 0.05 were represented.

2.16.6 Pseudotemporal ordering and construction of single-cell trajectory

In order to study the transcriptional dynamics during the process of leukaemia transformation, Monocle (v3.0) (Trapnell *et al.*, 2014) was used. To start with, a cell data set (CDS) object was created from raw data having expression matrix, cell metadata and gene annotations. After this, pre-processing was done by normalizing the data by log and size factor to address depth differences. Post normalization, PCA analysis was done in order to study the variance contributed by each PC (Fig 2.11).

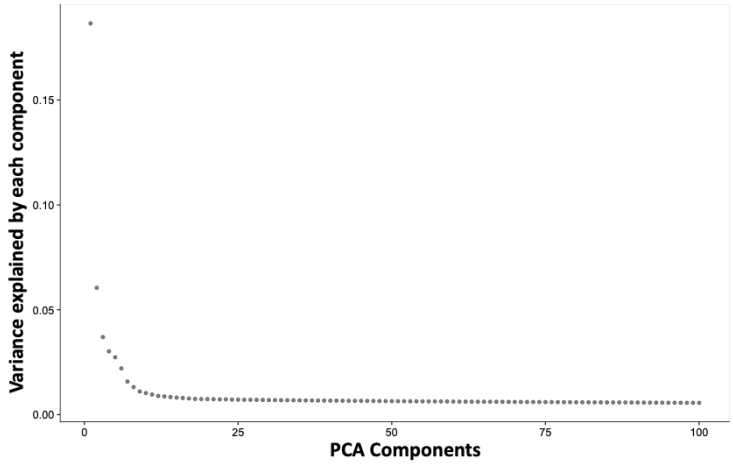


Figure 2.11: Pre-processing using Monocle v3.0.
Principal Component Analysis was done in order to observe the variance contributed by each PC and utilize this information to visualize cells in a reduced dimensional space.

To visualize the cells in a lower dimensional space, Uniform Manifold Approximation and Projection (UMAP) method was used which is integrated in Monocle 3. Further, cells were divided into larger well separated groups called partitions following Partition-based Graph Abstraction (PAGA) algorithm integrated in Monocle3 (Wolf *et al.*, 2019) and visualized on UMAP plot obtained from previous step. After grouping the cells, a principal graph was fitted for these partitions and further, each cell was ordered according to it's progress along the learned trajectory. This alignment of cells as a function of progress along the trajectory is termed 'pseudotime'.

For ordering the cells along pseudotime, the beginning of the biological process, termed as 'root' was defined by a region where most of the WT-2 months cells were present. For individual WT or *Kat2a* NULL cells trajectories, the root was defined by WT/ *Kat2a* NULL-2 months cells. After defining the root, each cell was ordered in pseudotime space with respect to the root.

The cellular compartments were identified by differential expression performed using DESeq2 as described above and further by performing Gene Ontology analysis using Panther.

2.16.7 Transcriptional variability measurement

For measuring transcriptional variability, pairwise distance method was used (Mohammed *et al.*, 2017). In order to measure this, top 500 variable genes from each subset under comparison were identified based on Distance to Median (DM) and correlation measures were calculated for these. These correlation measures were further converted into distance using equation below-

$$d = \sqrt{[(1 - \rho)/2]}$$

d = transcriptional variability
ρ = Spearman's correlation coefficient

The distance values representing transcriptional variability were plotted in the form of violin plot (Fig 2.12).

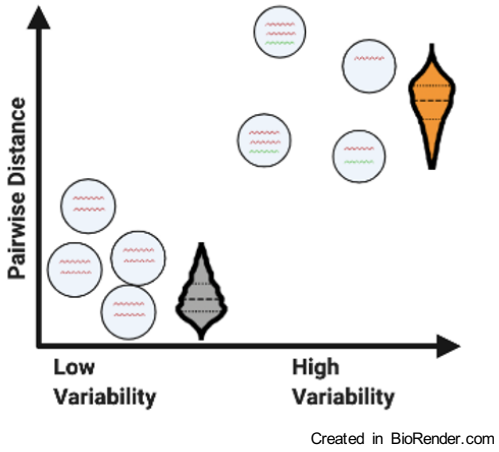


Figure 2.12: Transcriptional variability measurement.
Schematic representing heterogeneous transcription which can be measured by pairwise correlation method.

2.17 Single-cell ATAC sequencing

2.17.1 Sample preparation

The sample preparation (Fig 2.13) including building up of libraries for single-cell ATAC sequencing were done by members of Pina group prior to my joining. For this, the Kasumi-1 cells were taken and subjected to treatment with 100 µM MB-3 where 0.1% DMSO served as control. The cells were cultured the day before at $7*10^5$/ml, treated on D0 and collected 24 hours later (D1). The library preparation was done using C1 Single-Cell Auto Prep System (Fluidigm, Inc.) (Buenrostro *et al.*, 2015a).

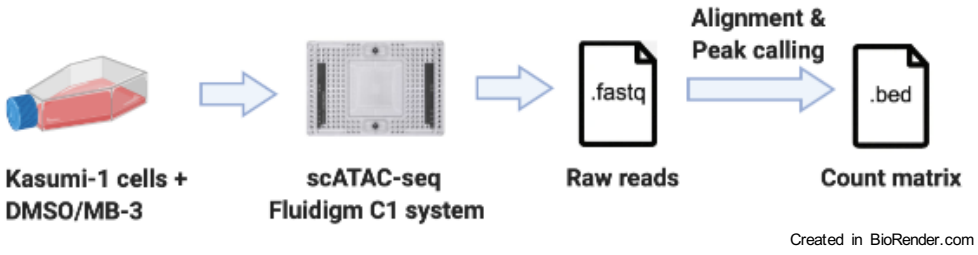


Figure 2.13: Sample preparation and generation of count matrix from scATAC-seq data. Schematic describing the different steps involved in generating scATAC-seq data, starting from the treatment of Kasumi-1 cells with KAT2A inhibitor MB-3 where DMSO served as control, these cells were then subjected to scATAC-seq library preparation using Fluidigm C1 system, the raw reads generated upon sequencing were then subjected to alignment and further utilised for peak calling in order to generate a workable matrix.

2.17.2 Matrix generation, pre-processing and filtering of data

The sequencing data generated was aligned using hg19 as reference genome and peak calling was done using MACS2 (Zhang *et al.*, 2008). This led to 74,284 peaks in DMSO and 111,898 peaks in MB-3 in the form of binary measure where 0 and 1 corresponds to the absence and the presence of a peak, respectively. These peaks were merged using Bedtools 2.27.0 (Quinlan and Hall, 2010) and filtered for the peaks which were present in at least 15 % of the cells. The strict filtering was done to avoid any false positive peaks and to balance the number of peaks obtained from higher number of MB-3 cells sequenced (50 vs. 38 in DMSO). Post-filtering, 4157 peaks were obtained in total which were subjected to differential accessibility analysis based on Fisher exact test and information gain as mentioned above. The p-values generated were corrected for the mean number of comparisons using Bonferroni (Haynes, 2013) and Benjamini Hochberg (Benjamini and Hochberg, 1995) correction method. The combined measure of frequency of peaks in DMSO and MB-3 cells and differential analysis resulted in 50 peaks which were characterized as DMSO unique, 520 peaks which were characterized as MB-3 unique and 3587 peaks which were attributed as common set of peaks.

2.17.3 Jaccard distance calculation

After obtaining the peak accessibility matrix, Jaccard distance method (Jaccard, 1901) was used to quantify the dissimilarity in which two cells vary in their peak accessibility. The Jaccard distance was calculated as the ratio between the number of peaks that are unique to a cell against all the peaks that are open in two cells.

2.17.4 Differential accessibility analysis

Differential accessibility analysis was done for a common subset of peaks in order to calculate statistical significance for quantitative changes in chromatin accessibility between DMSO and MB-3 treated cells. For this, two different approaches were followed- Information Gain and Fisher exact test.

Information gain is a measure of homogeneity and calculates the reduction in entropy based on which it suggested the peaks which differentiate MB-3 from DMSO. The information gain was calculated as-

$$Gain(P, P_{G_1}, P_{G_2}) = Entropy(P) - \sum_{v \epsilon \{G_1, G_2\}} \frac{|P_v|}{|P|} Entropy(P_v)$$

Where P is collection of all data, P_{G1} represents cells in G_1 and P_{G2} represents cells in G_2.

Fisher exact test was run on a peak-by-peak basis by organizing the open and closed (1's and 0's) for each peak in a 2 × 2 contingency table. The p-values obtained were then corrected for multiple comparisons using Bonferroni correction.

2.17.5 Dimensionality reduction analysis and k-medoid clustering

For easy visualization and downstream analysis, dimensionality reduction analysis was performed using tSNE. The cells were further clustered using k-medoid algorithm which breaks the dataset into different partitions and attempts to minimize the distance between points assigned in a cluster and a point designated as the centre of that cluster.

2.17.6 Genomic Regions Enrichment of Annotations Tool (GREAT)

In order to model the regulatory landscape corresponding to DMSO and/or MB-3 specific peaks, GREAT analysis (Mclean *et al.*, 2010) was done. The tool calculated region-gene associations for each subset of peaks with respect to hg19 species assembly.

2.17.7 Annotation of peaks

Annotation of different subset of peaks was done using HOMER (Duttke *et al.*, 2019). The peaks were provided in the form of Browser Extensible Data (BED) format having information about chromosome, starting/ending position and strand (+/-) specificity. The peaks were annotated using hg19 as reference genome.

2.17.8 Transcriptional variability calculation

Transcriptional variability measurement was done using pairwise distance method mentioned above. For this, the peaks which were found to have higher accessibility in MB-3-I cluster as compared to DMSO or MB3-II were annotated for closest genes using HOMER as mentioned above. In order to measure transcriptional variability using scRNA-seq from *RUNX1-RUNX1T1(9a)* pre-leukaemia, the genes obtained post annotation were converted to mouse homolog using BioMart (Smedley *et al.*, 2009).

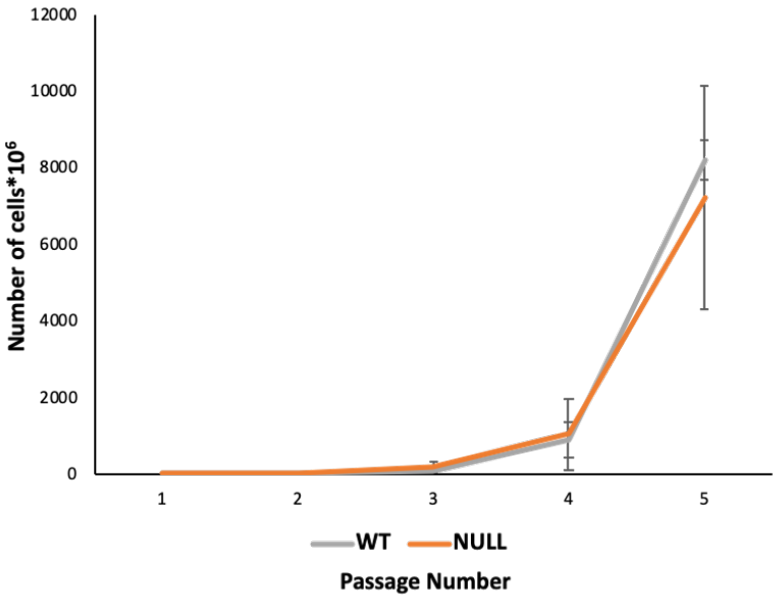
Pairwise correlation was calculated using top 500 variable genes for each subset, calculated as mentioned above.

2.18 Generation of *MLL-AF9* primary cell lines

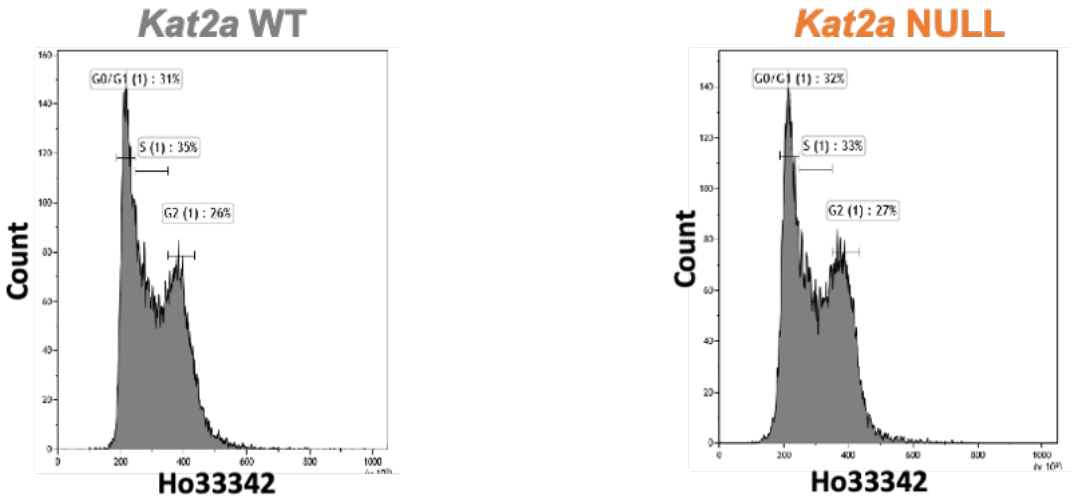
MLL-AF9 primary cultures were set-up using *Kat2a* WT and *Kat2a* NULL BM cells obtained from mice and transduced with *MLL-AF9 in vitro* as described above. The cells obtained were enriched for transformants by three rounds of serial re-plating in MethoCult M3434. These transformed cells were then cultured in R20 supplemented with 20ng/ml each of mSCF, mIL-3, mIL-6 at a density of $0.2*10^6$ cells/ ml and passaged when they reached a density of $1*10^6$cells/ ml.

The primary cell lines generated *in vitro* were then carefully observed to study any differences in growth patterns (Fig 2.18A). Live and dead cell counting was performed every alternate day for both genotypes.

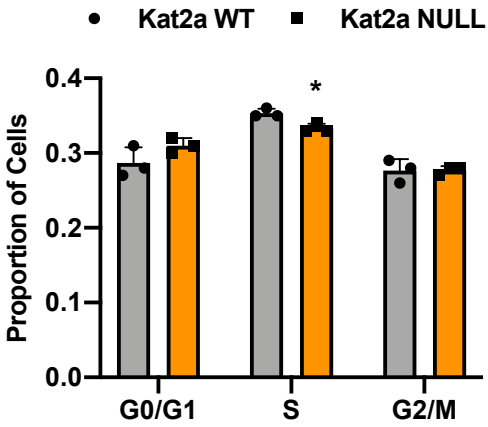
A


12000
10000
8000
6000
4000
2000
0
Number of cells*10^6
1
2
3
4
5
WT
NULL
Passage Number

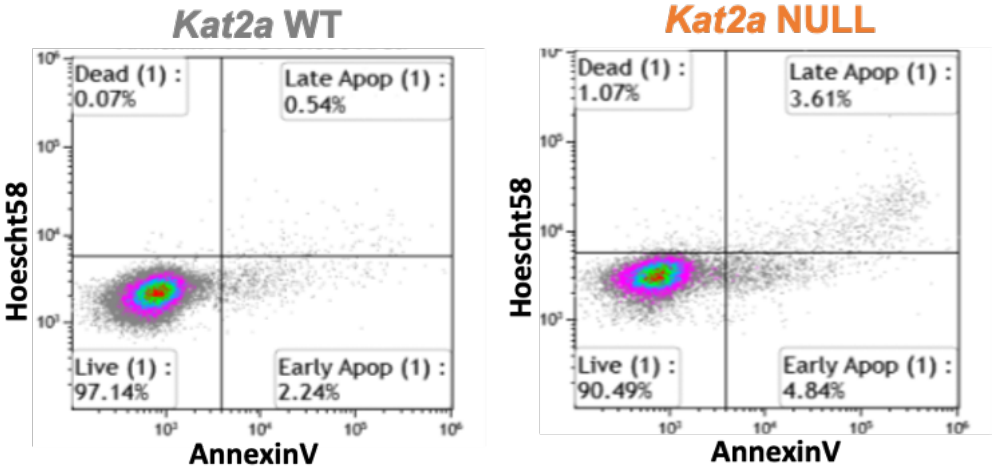
B


Kat2a WT
G0/G1 (1) : 31%
S (1) : 35%
G2 (1) : 26%
Count
Ho33342
Kat2a NULL
G0/G1 (1) : 32%
S (1) : 33%
G2 (1) : 27%
Count
Ho33342

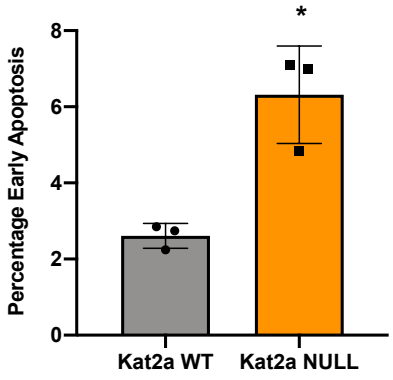
C


Kat2a WT
Kat2a NULL
Proportion of Cells
0.4
0.3
0.2
0.1
0.0
*
G0/G1
S
G2/M

D


Kat2a WT
Kat2a NULL
Hoescht58
AnnexinV
Dead (1) : 0.07%
Late Apop (1) : 0.54%
Live (1) : 97.14%
Early Apop (1) : 2.24%
Dead (1) : 1.07%
Late Apop (1) : 3.61%
Live (1) : 90.49%
Early Apop (1) : 4.84%

E


Percentage Early Apoptosis
8
6
4
2
0
*
Kat2a WT
Kat2a NULL

F

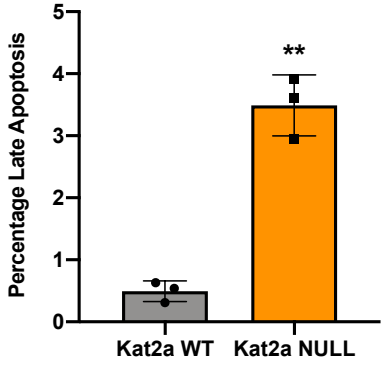

Percentage Late Apoptosis
5
4
3
2
1
0
**
Kat2a WT
Kat2a NULL

Figure 2.14: Characterization of *MLL-AF9* primary cell lines.
(A) Growth curve representing number of cells (per $1*10^6$) for both *Kat2a* WT and *Kat2a* NULL at each passage (mean ± SD), (**B)** Cell cycle profile obtained for both genotypes using flow cytometry analysis for Hoescht 33342 representing G0/G1, S and and G2/M phase of cell cycle, (**C)** Proportion of cells in each phase of cell cycle for both genotypes (n=3/genotype, mean ± SD, Student's t-test $p<0.05$*), (**D)** Flow cytometry plots representing early and late apoptosis marked by Annexin V and Hoescht58 for both genotypes, € Representative plot for early apoptotic population of cells for both genotypes (n=3/genotype, mean ± SD, Student's t-test $p<0.05$ *), (**F)** Representative plot for late apoptotic population of cells for both genotypes (n=3/genotype, mean ± SD, Student's t-test $p<0.01$**).

Post three passages, cell cycle analysis was done to see any differences during different phases of cell cycle (Fig 2.18B). For this, $1*10^6$ cells of both *Kat2a* WT and *Kat2a* NULL cells were incubated in R20 medium with Hoescht33342 (2µg/ml) at 37°C for 2 hours and cells were acquired on Gallios Flow Cytometer (Beckman Coulter) and data was analysed using Kaluza software suggesting a reduction in S phase in *Kat2a* NULL cells (Fig 2.18C).

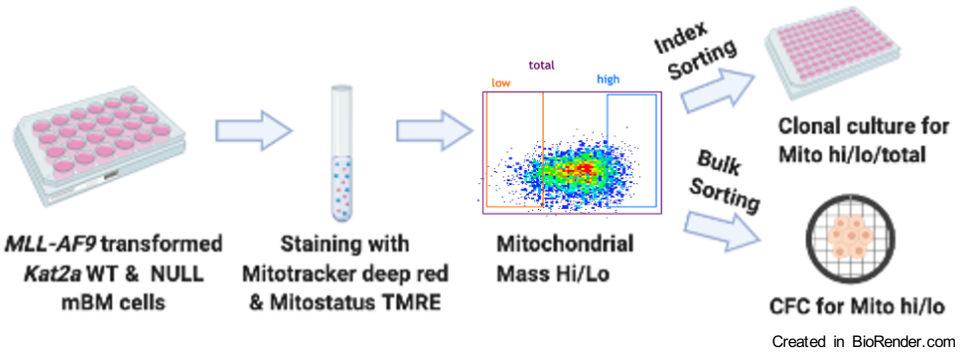
Apoptosis studies were also done in parallel for both *Kat2a* WT and *Kat2a* NULL cells (Fig 2.18D). The cells were incubated in 5µl of Annexin V APC (BioLegend) for 15 minutes and resuspended in Hoescht33258 (1:10000). Acquisition was done on Gallios Flow Cytometer (Beckman Coulter) and data was analysed using Kaluza software suggesting an increase in both early and late apoptosis in *Kat2a* NULL cells (Fig 2.18E, F).

2.19 Mitochondrial analysis

For mitochondrial analysis, 10^5 cells/ sample were resuspended in 100µl of R20 medium supplemented with 10ng/ µl mIL-3, 10ng/ µl mIL-6 and 20ng/ µl mSCF as mentioned above. Unstained and single colour controls were prepared from a pool of all samples with 10^5 cells/ control and resuspended in same media. For each sample and control, 1µl of Mitotracker- Deep Red and Mitostatus- Tetramethylrhodamine ethyl ester (TMRE) were added to obtain a final concentration of 500nM and 100nM respectively. The addition of dyes was performed in dark. The cells were incubated for 30 minutes at 37°C in CO_2 incubator. Post incubation, 700µl of R20 was added and the cells were centrifuged at 2000 rpm for 5 minutes. The supernatant was removed, and samples were resuspended in 200µl of R20 supplemented with cytokines as

mentioned above and Hoescht58 at a dilution of 1:10000. The samples were acquired on Gallios Flow Cytometer (Beckman Coulter) at low/ medium speed, gating was done on GFP^+ for *RUNX1-RUNX1T1(9a)* and YFP^+ for *MLL-AF9* transformed cells and data was analysed using Kaluza software.

A



B

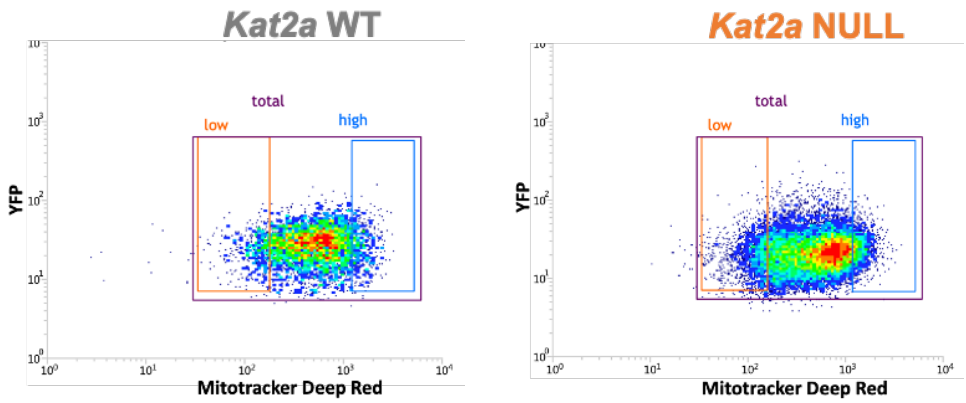


Figure 2.15: Mitochondrial analysis and gating strategy.

(A) Experimental strategy for Mito hi/Mito lo analysis where *MLL-AF9* transformed cells were stained with mitochondrial dyes and sorted as Mito hi and Mito lo, **(B)** Gating strategy for defining low and high population for WT and *Kat2a* NULL cells transformed with *MLL-AF9.*

In order to study the phenotypic differences between high mitochondrial mass (Mito hi) and low mitochondrial mass (Mito lo), *MLL-AF9* mBM transformed cells post staining with Mitostatus and Mitotracker were sorted based on mitochondrial mass high/ low (Fig 2.19A, 2.19B). Index sorting was done for both *Kat2a* WT and *Kat2a* NULL where single cell was collected in each well of a 96-well plate from Mito hi, Mito lo and Mito total fractions in 200 µl of R20 medium supplemented with cytokines. These single cell clonal cultures were kept at

37°C in CO_2 incubator and tracked for cell counts in each well for 7 days in order to visualize any differences in clonal expansion abilities of these cells.

The cells were also sorted in bulk in order to study any differences in self-renewal potential of the mitochondrial fractions in both genotypes. The sorted cells ~2000 cells/ sample were maintained in the form of CFC assays in technical duplicates as mentioned above. The colonies obtained were scored after 5 days.

2.20 Tigecycline inhibition

To study whether inhibition of mitochondrial translation activity in *Kat2a* WT cells phenocopies that of *Kat2a* NULL cells, *Kat2a* WT BM cells transformed with *MLL-AF9* were taken and treated with 2.5µM and 5µM of Tigecycline where R20 medium served as control. 3.3µl of Tigecycline (equivalent to respective concentrations) or medium was added directly to MethoCult M3434 and vortexed thoroughly. 10,000 cells/condition were added to the medium and plated in duplicates. The colonies obtained were scored after 5-7 days of plating and analysed for cell surface marker, Cd117/ c-Kit using Gallios Flow Cytometer (Beckman Coulter). The data obtained was analysed using Kaluza software.

Further, to study the effect of mitochondrial translation inhibition during leukaemia transformation, 50,000 cells each of *Kat2a* WT BM cells transduced with *RUNX1-RUNX1T1(9a)* were treated with 2.5µM of Tigecycline (Tocris) where R20 medium served as control. Colonies were scored after 7-10 days of plating and further re-plated with 10,000 cells/condition for plate 2 and 4,000 cells/condition for further platings in presence of Tigecycline. The colonies obtained at plate 5 were analysed using flow cytometry following the procedure described above for c-Kit (Fig 2.16).

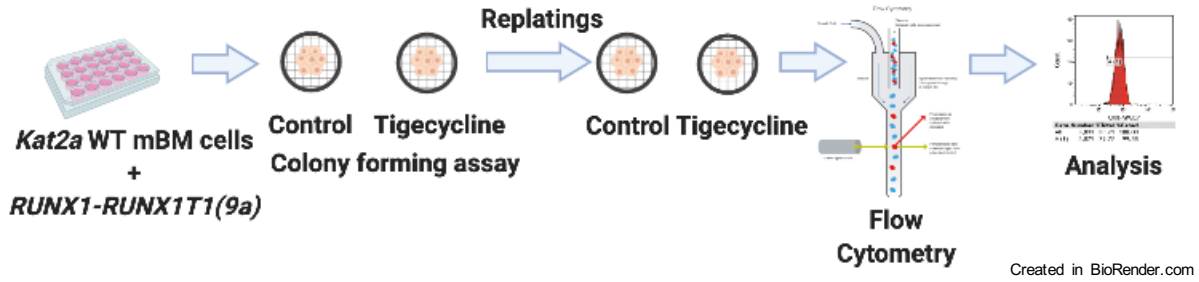


Figure 2.16: Schematic for Tigecycline experiment.

Experimental strategy for studying the effect of mitochondrial translation inhibition upon *RUNX1-RUNX1T1(9a)* transformation using Tigecycline and performing flow cytometry analysis.

2.21 O-propargyl-puromycin (OP-Puro) assay

OP-Puro assay was done in order to study differences in protein synthesis. For this, *Kat2a* WT and *Kat2a* NULL cells transformed with *RUNX1-RUNX1T1(9a)* or having *Idh1*R132H, 20-weeks post *pIpC* were serially re-plated during colony forming assays as described above. At each plating, $1*10^6$ cells each of *Kat2a* WT and NULL cells were taken where $0.8*10^6$ cells were cultured in R20 medium supplemented with 20ng/ml of mSCF, 10ng/ml of mIL-3 and 10ng/ml of mIL-6 having 12.5 µM OP-Puro (Thermo Scientific). Rest $0.2*10^6$ cells were cultured in the same medium where OP-Puro was replaced with PBS. The cells were kept in culture for 1 hour in CO_2 incubator. Post incubation, cells were washed, resuspended in PBE and surface staining was done with c-Kit-APCC7, Sca1-PC7, CD11b-Biotin, Gr1-Biotin, Streptavidin Brilliant Violet 605 (all BioLegend) as described above. Respective single colour controls were prepared with a pool of treated and non-treated cells. Post staining, each sample was resuspended in 500µl of 1% Paraformaldehyde (PFA), vortexed thoroughly and incubated for 20 min at 4°C for fixation. The cells were washed twice with PBS post fixation and resuspended in 200µl of cold permeabilization buffer (PBS +3% FBS +0.1% saponin) for 5 minutes in dark. Post incubation, samples were centrifuged, supernatant was removed and immediately proceeded for Azide-Alkyne Cycloaddition reaction. Each sample was resuspended in 500µl of following reaction mixture-

440µl 1x Reaction Buffer (component A; Click-iT Cell Reaction Buffer Kit, Life Technologies)

10µl $CuSO_4$ (component B)

1µl Alexa Fluor 647-Azide (AF647-Azide) (Life Technologies)

50µl Additive (component C)

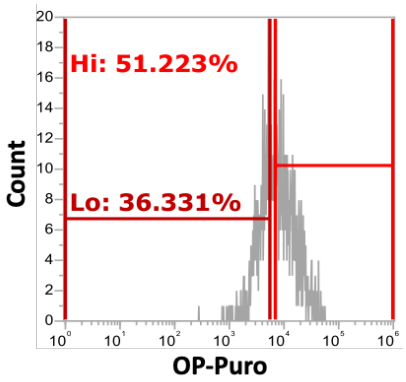


Figure 2.17: Gating strategy for OP-Puro analysis.

Gating strategy differentiating OP-Puro low (Lo) and high (Hi) population based on unstained and single colour controls for both *RUNX1-RUNX1T1(9a)* or *Idh1*R132H.

The samples were thoroughly vortexed and incubated for 30 minutes in dark at room temperature. Post incubation, samples were washed twice with 1ml of ice cold permeabilization buffer and resuspended in 200ul of PBS. All centrifugation steps were done at 2000 rpm for 5 minutes at 4°C. The samples were acquired on Attune NxT Flow Cytometer (Thermo Scientific) and analysed using Attune Nxt Software version 3.2.1. Gating strategy was based on unstained and single colour control defining OP-Puro low and high populations (Fig 2.21).

2.22 S6K1 inhibition

Functional studies on protein synthesis ablation were done using p70 ribosomal S6 kinase (S6K1) inhibitor (Tocris). For this, *Kat2a* WT BM cells transduced with *RUNX1-RUNX1T1(9a)* or having *Idh1*R132H, 4-weeks post *pIpC*, were treated with either 3.3 µl of control DMSO or S6K1 inhibitor (equivalent to final concentration of 10µM). The inhibitor or DMSO was directly added to MethoCult M3434 and vortexed thoroughly to which cell suspension with 10,000 cells/ condition were added and plated. The colonies were scored after

7-10 days and single cell suspensions obtained were subjected to OP-Puro analysis similarly as mentioned above (Fig 2.22). 10,000 cells/ condition were re-plated and the same procedure was repeated until third plating.

Figure 2.18: Experimental strategy for S6K1 inhibition experiment.
Kat2a WT cells transformed with *RUNX1-RUNX1T1(9a)* or *Idh1R132H* were treated with S6K1 inhibitor with DMSO as vehicle. Colony scoring and OP-Puro analysis was done at each plating.

2.23 Statistical Analysis

All experiments were done in triplicates except where mentioned. The data is plotted as Mean ± Standard deviation and Student's t-test was performed using Graphpad Prism 8.0 software. In case of scRNA-seq and scATAC-seq analysis, adjusted p-value was calculated using Bonferroni's method of correction.

3 Functional characterization of *RUNX1-RUNX1T1(9a)* and *Idh1*R132H leukaemia in a *Kat2a* knockout genetic background

In this chapter, I studied the functional role of *Kat2a*, an established regulator of cell-to-cell transcriptional variability, in a disease initiation set-up. Herein, I describe the experiments which were conducted in order to study the impact of *Kat2a* loss in 2 mouse models of AML, namely, the ones initiated by *RUNX1-RUNX1T1(9a)* and *Idh1*R132H. These models typically require additional mutations for leukaemia progression and thus, represent forms of human disease with a prolonged pre-leukaemia phase. In order to study the role of *Kat2a* in these disease models, I made use of *Kat2a* conditional knockout mice model, available in the lab, (MGI:3801321) (Lin *et al.*, 2008) with interferon response-inducible *Mx1-Cre*$^{+/-}$ (Kühn *et al.*, 1995a) in a C57Bl/6 background. The method utilizes interferon γ (IFNγ)-inducible Mx dynamin-like GTP-ase 1 (Mx1) promoter which controls Cre recombinase expression (Kühn *et al.*, 1995b), activation of which allows targeted silencing of *Kat2a in vivo* in haematopoietic system. *Kat2a* locus excision was obtained through the treatment of experimental (*Kat2a* excised- *Kat2a* NULL or *Kat2a* knockout) and control (*Kat2a* floxed- *Kat2a* WT) mice of both genotypes with intra-peritoneal polyinosylic-polycytidylic (pIpC) acid (Chan *et al.*, 2011). My lab has confirmed previously that *Kat2a* deletion does not perturb normal haematopoiesis and thus, preserves candidate progenitor cells-of-origin for leukaemia transformation (Domingues *et al.*, 2020). In line with this, I studied *RUNX1-RUNX1T1(9a)* pre-leukaemia by making use of bone marrow (BM) cells retrovirally transduced with *RUNX1-RUNX1T1(9a)* and injected back into C57/BL6 mice. For studying *Idh1*R132H pre-leukaemia, I crossed *Idh1*R132H conditional knock-in model (obtained from a collaborator as described in Methods) with *Kat2a*$^{fl/fl}$ mouse in order to generate a conditional *Idh1*R132H *Kat2a* fl/fl mouse model which was further utilized for pre-leukaemia studies. The pre-leukaemia cells were collected at different time points post *RUNX1-RUNX1T1(9a)* transduction or *Idh1*R132H mutation activation and were phenotypically characterized using flow cytometry analysis, colony forming replating assay, and peripheral blood sampling during the course of pre-leukaemia progression, as described in detail in this chapter.

3.1 Loss of *Kat2a* promotes pre-leukaemia to leukaemia acceleration in *RUNX1-RUNX1T1(9a)* model of leukaemia

To study the loss of *Kat2a* in *RUNX1-RUNX1T1(9a)* model, BM cells were extracted from *Kat2a* WT and *Kat2a* NULL animals post 4-6 weeks of *pIpC* treatment (Methods). The BM cells obtained were enriched for lineage depleted cells ($B220^-$, $Ter119^-$, $Cd11b^-$, $Gr1^-$ and $Cd3e^-$) and transduced with *RUNX1-RUNX1T1(9a)* overexpressing plasmid *in vitro.* These transduced BM cells were injected into lethally irradiated C57/BL6 animals as described in methods in order to study differences in leukaemia progression *in vivo* upon *Kat2a* loss (Fig 3.1A). There were 17 animals injected for BM cells from each genotype, of which 5 animals per genotype were utilized for pre-leukaemia studies. The animals were routinely observed for the presence of clinical symptoms of hunched posture, inappetence, and lethargy, indicative of leukaemia development. The first *Kat2a* NULL animal died of leukaemia post 104 days of transplantation. In contrast, the first *Kat2a* WT animal to develop leukaemia did so post 171 days of transplantation. Overall, an acceleration in leukaemia progression was observed in *Kat2a* NULL animals compared to *Kat2a* WT, suggesting that loss of *Kat2a* promotes *RUNX1-RUNX1T1(9a)* leukaemia progression (Fig 3.1B). The analysis includes animals having developed leukaemia until 405 days of transplantation which is when the experiment was terminated. The animals which died of reasons other than leukaemia were excluded from the statistical analysis.

Animals that developed leukaemia were inspected for the presence of *RUNX1-RUNX1T1(9a)* transformed cells. For this, the animals were culled using schedule 1 method of euthanasia; BM cells and spleen cells were isolated and stained with a cocktail of antibodies (Methods). Flow cytometry analysis suggested that 30-60% of the total population was GFP^+ indicating *RUNX1-RUNX1T1(9a)* mediated transformation (Fig 3.1C and 3.1D). A significant proportion of these transformed cells were c-Kit^+ indicating the presence of early progenitor cells in leukaemic animals irrespective of genotypes (Fig 3.1C and 3.1D). The gating set-up was done on the basis of unstained and single colour controls for the individual experiment.

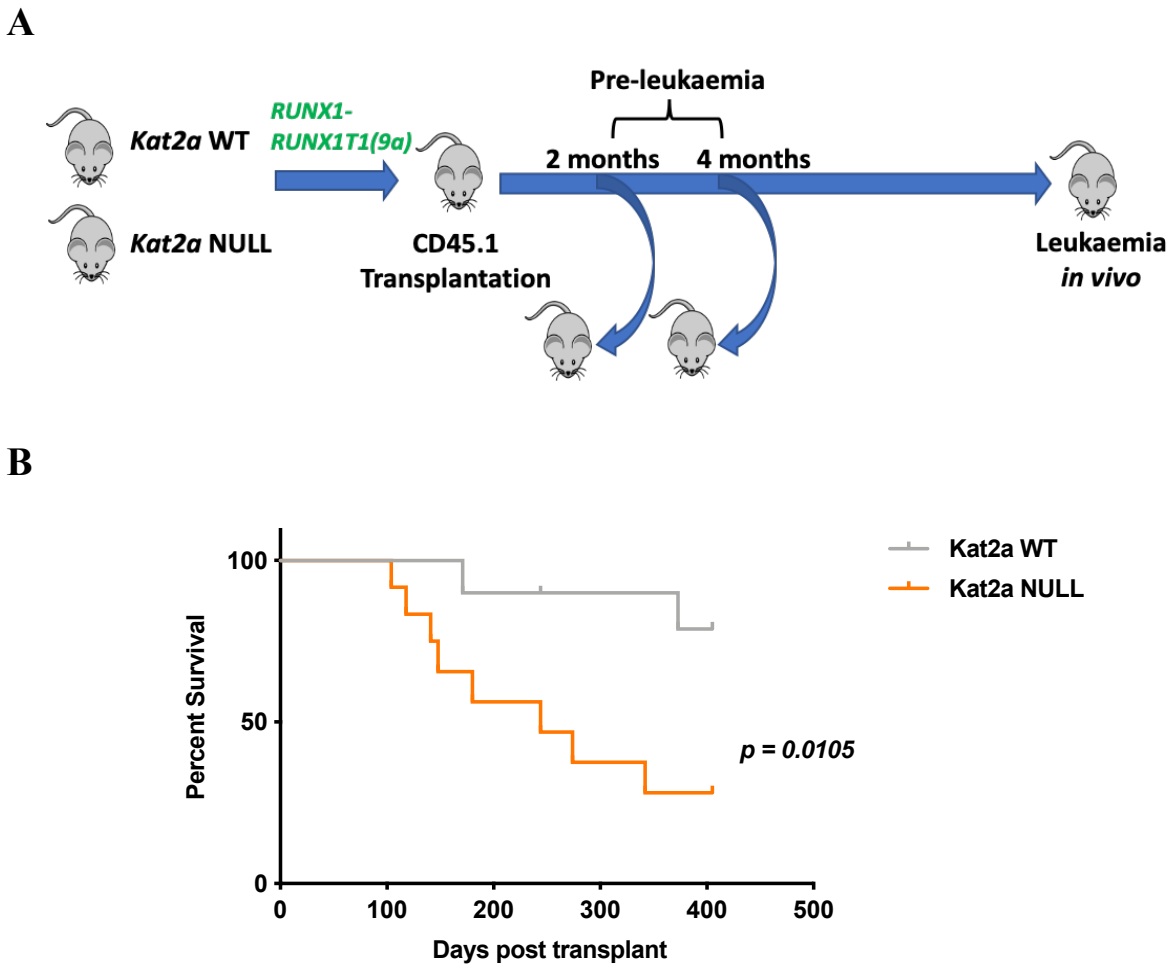

A
Kat2a WT
Kat2a NULL
RUNX1-
RUNX1T1(9a)
CD45.1
Transplantation
Pre-leukaemia
2 months
4 months
Leukaemia
in vivo
B
Percent Survival
100
50
0
0
100
200
300
400
500
Days post transplant
p = 0.0105
Kat2a WT
Kat2a NULL
C
Kat2a WT
GFP
C-Kit
GFP+ (1) : 62%
D
Kat2a NULL
GFP
C-Kit
GFP+ (1) : 42%

Figure 3.1: Functional characterization of *RUNX1-RUNX1T1(9a)* leukaemia.
(A) Schematic diagram representing experimental set-up. Bone marrow cells from *Kat2a* WT and *Kat2a* NULL animals were transformed with *RUNX1-RUNX1T1(9a)* overexpression plasmid having GFP reporter and transplanted into lethally irradiated CD45.1 animals (n=17 animals/genotype). Pre-leukaemia studies were done 2 months (n= 3/genotype) and 4 months (n=2/genotype) post-transplantation, **(B)** Survival curve representing percentage survival with respect to the number of days post transplantation for CD45.1 animals transplanted with *Kat2a* WT and *Kat2a* NULL cells (n=12 animals/genotype, Log rank test, p=0.0105), **(C)** Representative flow cytometry analysis of bone marrow cells obtained from one of the *Kat2a* WT animals which developed leukaemia. The plot highlights GFP^+ cells indicating *RUNX1-RUNX1T1(9a)* transformation, **(D)** Representative flow cytometry analysis of bone marrow cells obtained from for one of the *Kat2a* NULL animals which developed leukaemia.

The leukaemia studies were complemented by terminal peripheral blood analysis to understand differences in haematological parameters in *Kat2a* NULL animals over the period of leukaemia development (Fig 3.2). For this, blood sampling was done prior to culling for all the animals, irrespective of leukaemia induced death and samples were analysed using Vet abc automated counter (Methods). *Kat2a* NULL and *Kat2a* WT samples displayed no significant differences in any of the haematological parameters including WBC (Fig 3.2A), RBC (Fig 3.2B), and Haemoglobin (Fig 3.2C). Further, liver and spleen size and weight were measured between the genotypes to assess infiltration of the disease. There was no significant difference observed when comparing liver weight between the genotypes (Fig 3.2D left). Similar results were observed when the analysis was restricted to the animals which developed leukaemia (Fig 3.2D right). Similarly, no significant differences were observed for spleen size upon loss of *Kat2a* (Fig 3.2E left). Again, similar results were observed when the analysis was restricted to animals which developed leukaemia (Fig 3.2E right).

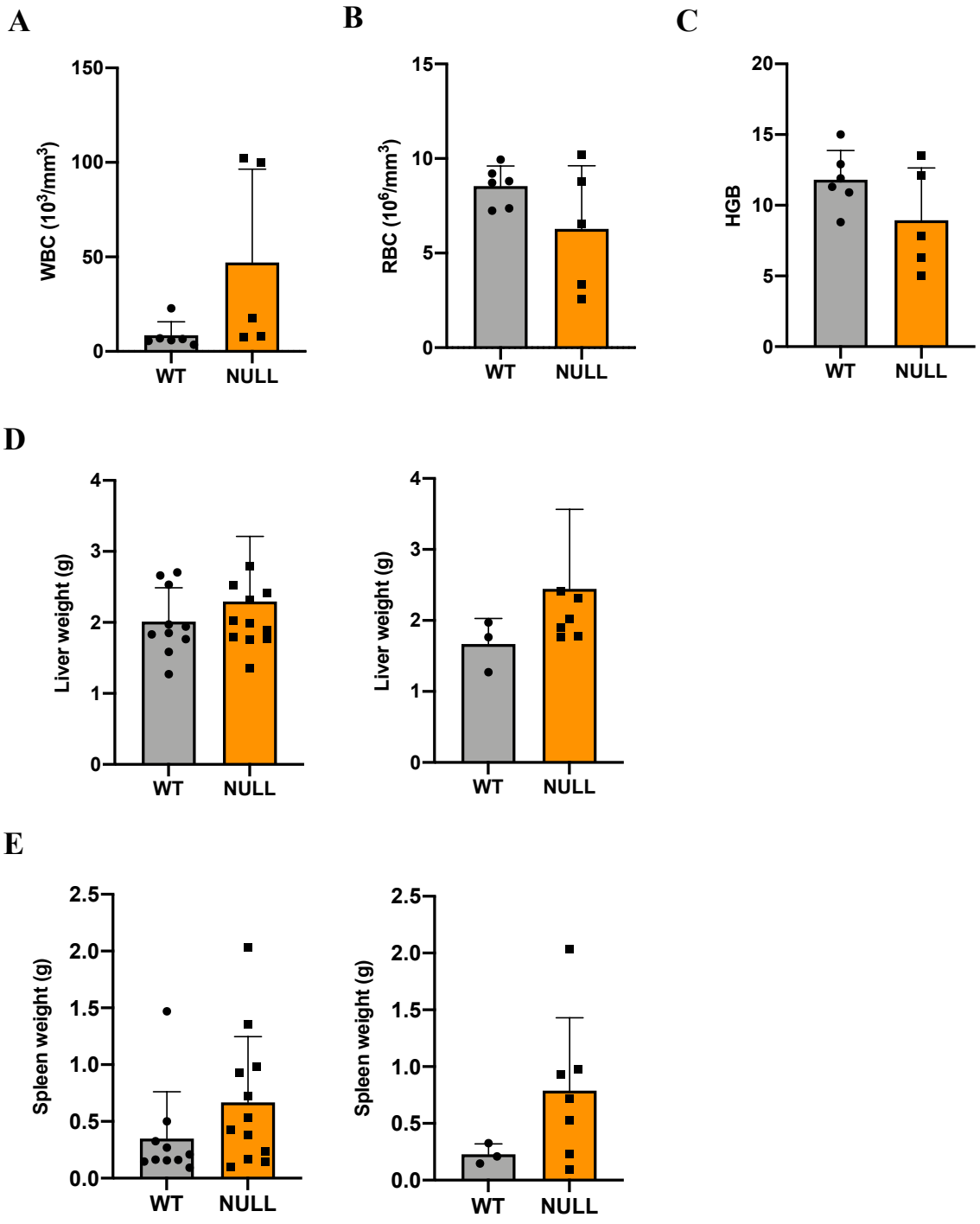

A
WBC (10^3/mm^3)
150
100
50
0
WT
NULL
B
RBC (10^6/mm^3)
15
10
5
0
WT
NULL
C
HGB
20
15
10
5
0
WT
NULL
D
Liver weight (g)
4
3
2
1
0
WT
NULL
Liver weight (g)
4
3
2
1
0
WT
NULL
E
Spleen weight (g)
2.5
2.0
1.5
1.0
0.5
0.0
WT
NULL
Spleen weight (g)
2.5
2.0
1.5
1.0
0.5
0.0
WT
NULL

Figure 3.2: Peripheral blood analysis and terminal leukaemia burden study.
Peripheral blood analysis to study different haematological parameters including, **(A)** White Blood Cells, **(B)** Red Blood Cells, **(C)** Haemoglobin, for both *Kat2a* WT and *Kat2a* NULL animals (plots represent analysis for n=6 animals for *Kat2a* WT and n=5 animals for *Kat2a* NULL, mean ± SD, Student's t-test) **(D)** Comparison of liver weight between *Kat2a* WT and *Kat2a* NULL animals (left) (plots represent analysis for n=10 animals for *Kat2a* WT and n=12 animals for *Kat2a* NULL, mean ± SD), Comparison of liver weight between *Kat2a* WT and *Kat2a* NULL animals which developed leukaemia (plots represent analysis for n=3 animals for *Kat2a* WT and n=7 animals for *Kat2a* NULL, mean ± SD, Student's t-test) (right) **(E)** Comparison of spleen weight between *Kat2a* WT and *Kat2a* NULL animals (plots represent analysis for n=10 animals for *Kat2a* WT and n=12 animals for *Kat2a* NULL, mean ± SD, Student's t-test) (left), Comparison of spleen weight between *Kat2a* WT and *Kat2a* NULL animals which developed leukaemia (plots represent analysis for n=3 animals for *Kat2a* WT and n=7 animals for *Kat2a* NULL, mean ± SD, Student's t-test) (right).

3.2 *Kat2a* loss aids in the survival of *RUNX1-RUNX1T1(9a)* transformed cells at pre-leukaemia stage

To understand whether accelerated leukaemia progression upon loss of *Kat2a* is a consequence of advantageous transformation events at pre-leukaemic stages, peripheral blood sampling was done on a regular basis post transplantation in C57/BL6 mice injected with *RUNX1-RUNX1T1(9a)* transduced BM cells. Mononuclear cells extracted from peripheral blood samples were analysed for GFP^+ population which is representative of *RUNX1-RUNX1T1(9a)* transformed cells. Blood sampling was performed on 10 animals for each genotype starting from 5 weeks post transplantation until 48 weeks. *Kat2a* NULL animals showed an increase in *RUNX1-RUNX1T1(9a)* transformed cell population after 12 and 17 weeks of transplantation. This population remained stable at later stages of sampling suggesting a perpetuation of the transformed population which leads to accelerated leukaemia progression (Fig 3.3A, Two-way ANOVA analysis, p= 0.001***). On the other hand, *Kat2a* WT animals did not show any changes in accumulation of transformed cells suggesting the event is *Kat2a* specific.

In addition to the above, the collected blood samples were also analysed for various haematological parameters during disease progression. In general, there was a steady increase in WBCs until 17 weeks which remained stable until 48 weeks post transplantation (Fig 3.3B).

The proportion of HGB remained stable over the period of disease development (Fig 3.3C). The number of PLT were slightly reduced over the period, potentially due to aging (Fig 3.3D). However, there was a sharp reduction at 17 weeks which is likely due to presence of technical artifact since both the genotypes responded in similar manner. Overall, there were no genotype specific differences in haematological parameters over the period of disease progression, suggesting no leukaemia associated phenotypic changes upon *Kat2a* loss.

To assess whether these *in vitro* observations of fixation of *RUNX1-RUNX1T1(9a)* transformed cells replicated *in vivo*, *Kat2a* WT and *Kat2a* NULL animals were culled using schedule 1 method of euthanasia 2 months and 4 months post transplantation to study differences during pre-leukaemia transformation (n=3/genotype at 2 months and n=2/genotype at 4 months). The time points chosen reflected the *in vitro* fixation of the transformed population of cells. BM and spleen cells were isolated and stained with a cocktail of antibodies including CD117/ c-Kit (early progenitor marker), Cd11b/ Mac1 (monocyte/lin^+ marker), Sca1 (stem cell marker), Gr1 (granulocyte/ lin^+ marker/), CD34 (haematopoietic stem cell marker) and CD16/32/FCγR (myeloid progenitor marker) (Methods). The combination of markers was selected in order to define stem and progenitor markers in these *RUNX1-RUNX1T1(9a)* transformed BM cells obtained from pre-leukaemia animals (Weissman and Shizuru, 2008) (Doulatov *et al.*, 2012). The combined flow cytometry analysis for pre-leukaemia samples collected at both time points indicated a significant increase in the primitive population represented by GFP^+ Kit^+ $FC\gamma R^+$ indicative of myeloid progenitors in *Kat2a* NULL animals, (Fig 3.3E) in line with the *in vitro* observations obtained from peripheral blood sampling and suggesting perpetuation of *RUNX1-RUNX1T1(9a)* transformed pre-leukaemia clones.

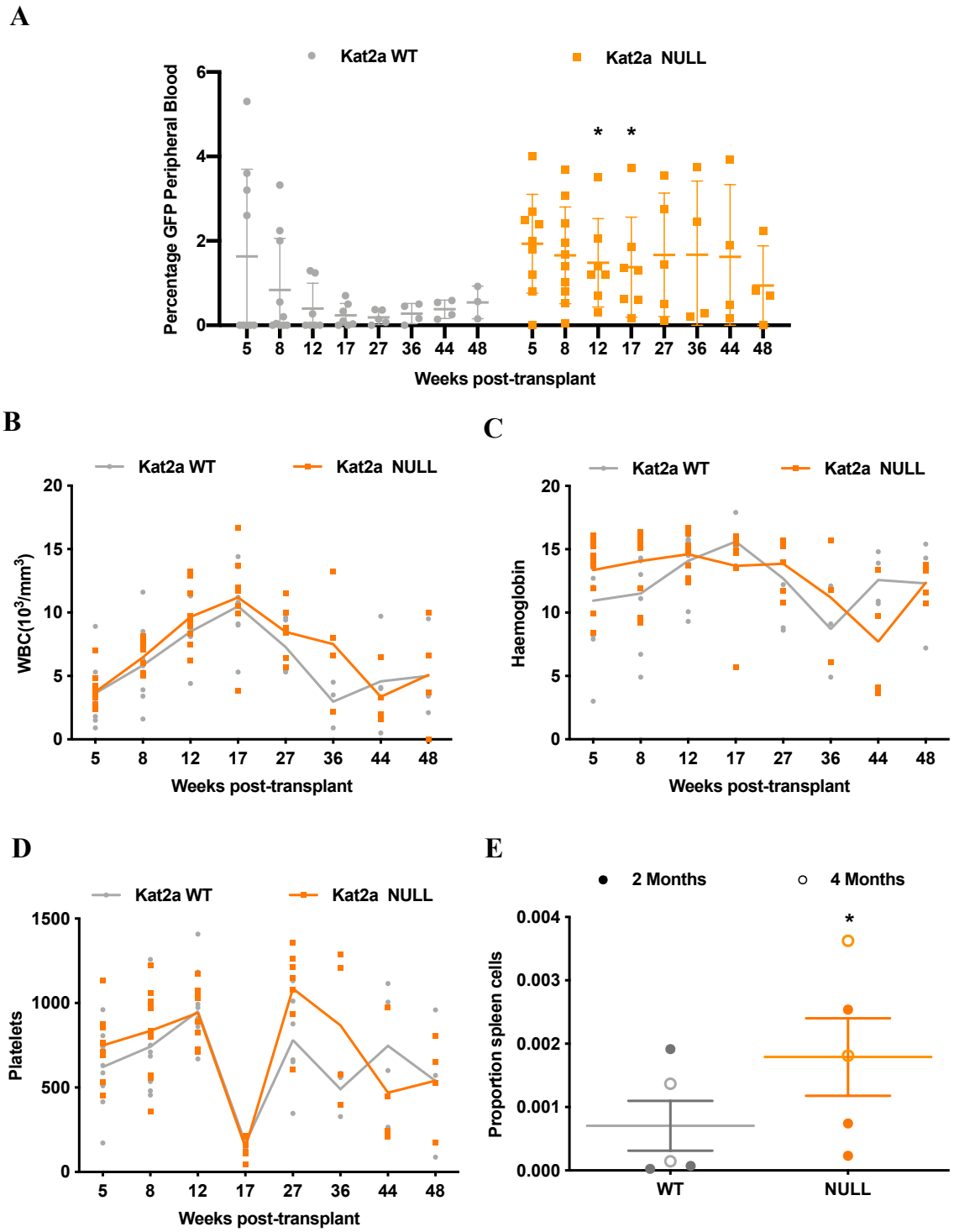

A
Kat2a WT
Kat2a NULL
Percentage GFP Peripheral Blood
Weeks post-transplant
B
WBC(10³/mm³)
C
Haemoglobin
D
Platelets
E
2 Months
4 Months
Proportion spleen cells
WT
NULL

Figure 3.3: ***In vivo*** **analysis of** ***RUNX1-RUNX1T1(9a)*** **pre-leukaemia.**

(A) Flow cytometry analysis representing GFP^+ population in mononuclear cells extracted from peripheral blood post 5, 8, 12, and 17 weeks of transplantation from both *Kat2a* WT and *Kat2a* NULL animals (n=10/genotype at 5 and 8 weeks, n=7/genotype at 12 and 17 weeks, mean ± SD, Student's t-test, p=0.0408* at 12 weeks and p=0.0438* at 17 weeks), Similar analysis to study any variations in haematological parameters between *Kat2a* WT and *Kat2a* NULL animals post 5, 8, 12, 17, 27, 36, 44 and 48 weeks of transplantation, **(B)** White blood cells, **(C)** Haemoglobin, **(D)** Platelets, (n=10 animals/genotype at 5, 8, 12 weeks, n=7 animals/genotype at 17 weeks, n=5 animals/genotype from 27 weeks onwards) **(E)** Flow cytometry analysis for pre-leukaemia animals processed post 2 and 4 months of transplantation. The plot represents leukaemic engraftment as defined by GFP^+ $c\text{-}Kit^+$ $FC\gamma R^+$ population in spleen cells from pre-leukaemia animals (n=3 animals/genotype for 2 months and n=2 animals/genotype for 4 months, mean ± SD, Student's t-test, p=0.0467*).

3.3 ***RUNX1-RUNX1T1(9a)*** **transformed cells show an increase in self-renewal capacity upon loss of** ***Kat2a***

After confirming perpetuation of *RUNX1-RUNX1T1(9a)* transformed cells in *Kat2a* NULL animals during pre-leukaemia, I wanted to study whether there are any differences in self-renewal capacity of these cells during the process of transformation *in vivo*. To assess this, serial re-plating colony assay was performed. I started with BM cells obtained from both *Kat2a* WT and *Kat2a* NULL (n=3/genotype) at 2 months as mentioned above. Initial plating was done with 50,000 cells/sample in technical duplicates and the colonies were scored after 7-10 days of plating. Serial re-plating was done using 10,000 cells/condition until no colonies were discovered, which allowed studying the serial replating capacity, an indicator of self-renewal potential of *RUNX1-RUNX1T1(9a)* transformed cells. The colonies were categorized as Granulocyte (G), Erythroid (E), Macrophage (M), Granulocyte Macrophage (GM), and Granulocyte Erythroid Macrophage Megakaryocyte (GEMM), where GM and GEMM colonies indicate the significant expansion of transformants (Methods). There was no significant difference between the genotypes in terms of re-plating potential at 2 months (Fig 3.4A). However, a trend towards enhanced self-renewal in *Kat2a* NULL at plate 4 (n=1/genotype at plate 4) was observed. This indicated that while it is possible that *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells may have enhanced self-renewal potential, 2 months post transplantation may be too early to capture this. Therefore, study of pre-leukaemia

at a later time point is required. In line with this, similar analysis was performed for pre-leukaemia BM cells collected at 4 months (n=2/genotype done in technical duplicates). There was a trend towards enhanced self-renewal capacity of *Kat2a* NULL cells from plate 1 onwards and a significant increase at plate 3, indeed suggesting that loss of *Kat2a* promotes self-renewal potential of *RUNX1-RUNX1T1(9a)* transformed cells *in vivo* (Fig 3.4B).

To study whether a similar trend in enhanced self-renewal exists *in vitro*, an *in vitro* colony forming assay was performed prior to my joining in the lab. The experiment was conducted on *Kat2a* WT and *Kat2a* NULL BM cells (n=3/genotype) isolated from *Kat2a*$^{fl/fl}$ mice post 6-8 weeks of *Mx1-Cre* activation, which coincides with the time when stable locus excision is achieved based on previous laboratory observations. The cells were enriched for lineage depletion and transduced with *RUNX1-RUNX1T1(9a)* overexpression plasmid as described in methods. Post transduction, the cells were washed and maintained in colony forming assay at 4,000 cells/sample in technical duplicates and the colonies were scored after 7-10 days of plating. Serial re-plating was done again using 4,000 cells/condition until plating 5. Again, there was an increase in self-renewal capacity in *Kat2a* NULL at plate 4 and plate 5 in line with the *in vivo* experiments thus strengthening the finding that loss of *Kat2a* enhances colony forming potential of *RUNX1-RUNX1T1(9a)* transformed cells (Fig 3.4C).

After confirming enhanced self-renewal potential in *Kat2a* NULL cells during disease initiation, my lab members sought to understand if there were any changes in a maintenance set up. For this, *RUNX1-RUNX1T1(9a)* transduction was done on BM cells carrying conditional *Kat2a* floxed allele prior to Cre recombinase activation. Post transduction, cells were washed with media and 50,000 cells/condition were maintained in colony forming assay. After re-plating the obtained colonies twice, the cells were transduced with Cre recombinase expressing vector in order to generate *Kat2a* NULL phenotype or with empty vector (EV) which served as control. Post transduction, cells were maintained in colony forming set up where *Kat2a* NULL cells showed reduced colony forming potential as compared to control both at plate 1 and plate 2 (Fig 3.4D). This observation suggested that enhanced self-renewal capacity of *Kat2a* NULL cells observed *in vitro* and *in vivo* was a consequence of loss of *Kat2a* during early stages of leukaemia transformation. On the other hand, once the cells are transformed, not having *Kat2a* is disadvantageous for the re-plating potential of these cells.

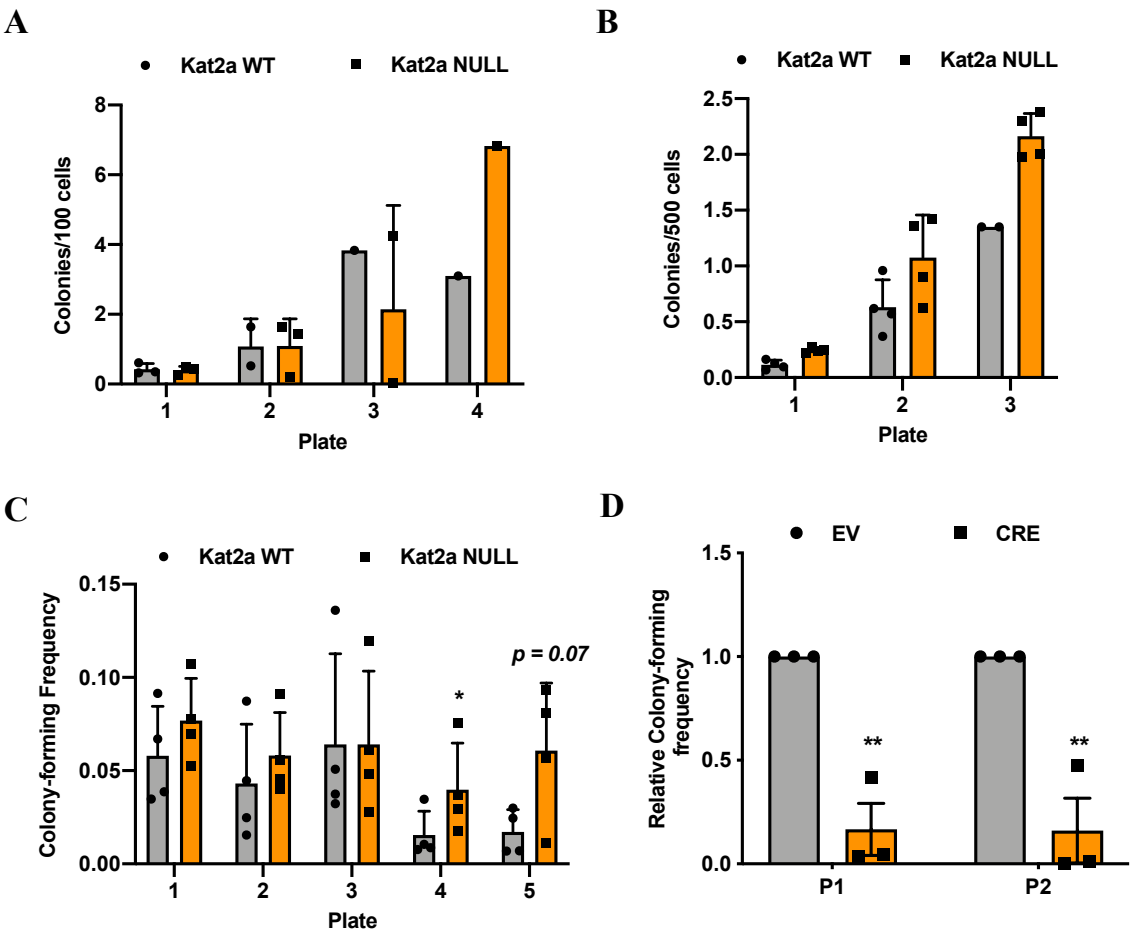


Figure 3.4: Colony forming unit analysis.

(A) Serial re-plating of colony forming cell assay from pre-leukaemia bone marrow cells obtained from *Kat2a* WT and *Kat2a* NULL animals post 2 months of transplantation (n=3 animals/genotype at P1, n=1 animal/genotype at P4, mean ± SD), **(B)** Serial re-plating of colony forming cell assay from pre-leukaemia bone marrow cells obtained from *Kat2a* WT and *Kat2a* NULL animals post 4 months of transplantation (n=2 animals/genotype at P1, n=1 animal for *Kat2a* WT and n=2 animals for *Kat2a* NULL at P4, mean ± SD), **(C)** Serial re-plating of colony forming cell assay from *in vitro* transformed *RUNX1-RUNX1T1(9a)* bone marrow cells obtained from *Kat2a* WT and *Kat2a* NULL cells (n=3 animals/genotype at each plating, mean ± SD, Student's t-test p=0.02* at P4 and p=0.07 at P5), **(D)** Serial re-plating of colony forming cell assay from *in vitro* transformed *RUNX1-RUNX1T1(9a)* bone marrow cells obtained from *Kat2a* WT where *Kat2a* floxed allele was obtained by *in vitro* transducing the cells with Cre recombinase expressing CRE plasmid where Empty vector (EV) served as control (n=3 animals/condition at each plating, mean ± SD, Student's t-test p=0.002** at P1 and p=0.005** at P2).

This was in line with the previous observation in the lab, where *MLL-AF9* transformed leukaemia cells, representative of a leukaemia maintenance model, showed enhanced differentiation upon loss of *Kat2a* (Domingues *et al.*, 2020).

3.4 *Kat2a* loss does not impact any haematopoietic compartment in *Idh1*R132H pre-leukaemia

Pre-leukaemia studies performed on *RUNX1-RUNX1T1(9a)* model in a conditional *Kat2a* genetic background suggested that loss of *Kat2a* aids in perpetuation of the transformed population of cells which have enhanced self-renewal capacity. To understand the general role of *Kat2a* in pre-leukaemia, it was important to understand whether these observations can be replicated in other pre-leukaemia models. For this, I studied *Idh1*R132H (*Idh1*mut) model developed in George Vassiliou's laboratory (Wellcome Trust Sanger Institute). *Idh1*R132H represents another model with a prolonged pre-leukaemia requiring incorporation of additional mutations (Introduction). The present model has a 30% leukaemia incidence 1-year post-mutation activation. To start, crossings were set up with *Kat2a* conditional knockout model in order to generate *Idh1*mut knock-in mice with conditional *Kat2a* allele dependent on *Mx1-Cre* activation. The activation of *Idh1* mutation and *Kat2a* excision were confirmed at different time points post *pIpC* injections as described in Methods.

In order to study the role of *Kat2a* in *Idh1*R132H leukaemia biology, it was important to investigate differences in various haematopoietic compartments in *Idh1*mut model upon *Kat2a* loss. For this, *Kat2a* HET and *Kat2a* NULL animals with *Idh1*mut were studied 4 weeks and 20 weeks post *pIpC* treatment (Fig 3.5A), reflecting pre-leukaemia development at two time points (n=3/genotype for each time point). The animals were sacrificed using schedule 1 method of euthanasia and were further utilized to study the cell populations belonging to different haematopoietic compartments. The combination of markers was selected in order to define stem and progenitor markers in *Idh1*R132H transformed BM cells obtained from pre-leukaemia animals (Weissman and Shizuru, 2008) (Doulatov *et al.*, 2012) (Sasaki *et al.*, 2012b). BM cells obtained were stained using a cocktail of antibodies including CD16/32-FITC, CD135-PE, CD117/C-Kit- APCCy7, Sca1-PECy7, CD34-APC as mentioned in Table B.4 (Annexure-B). Lineage exclusion cocktail included B220, Ter119, CD11b, Gr1, and CD3e

biotinylated antibodies along with streptavidin-conjugated Brilliant Violet 510 were also added as mentioned in Table B.4 (Annexure-B). The gating strategy for different cell populations including HSC, LMPP, MPP, GMP, CMP, MEP is described in methods. The experiments on these cells were performed by Oliwia Cyran, a placement student in the lab mentored by me whereas I performed the analysis.

Post 4 weeks of mutation activation, there were no significant differences between the two genotypes in any of the haematopoietic compartments including HSCs, suggesting that loss of *Kat2a* at early stages of leukaemia progression doesn't impact the BM cellularity (Fig 3.5B). Similarly, there were no significant changes post 20 weeks of mutation activation between the genotypes supporting the observation in 4 weeks samples (Fig 3.5C). However, between 4 weeks to 20 weeks, I observed an increasing trend in overall cellularity, in particular for the CMP and GMP population. This trend was observed irrespective of genotype but was not statistically significant. Further, I looked at KSL and KL populations where no significant differences were found between the genotypes at 4 weeks (Fig 3.5D) and 20 weeks (Fig 3.5E), however, there was a slight increasing trend in KSL in *Kat2a* NULL cells. While comparing these populations from 4 weeks to 20 weeks, there was an increasing trend in KL at 20 weeks indicating putative pre-leukaemia progression. Overall, loss of *Kat2a* did not impact any haematopoietic compartment in *Idh1*R132h leukaemia model, consistent with previous observations in our lab (Domingues *et al.*, 2020).

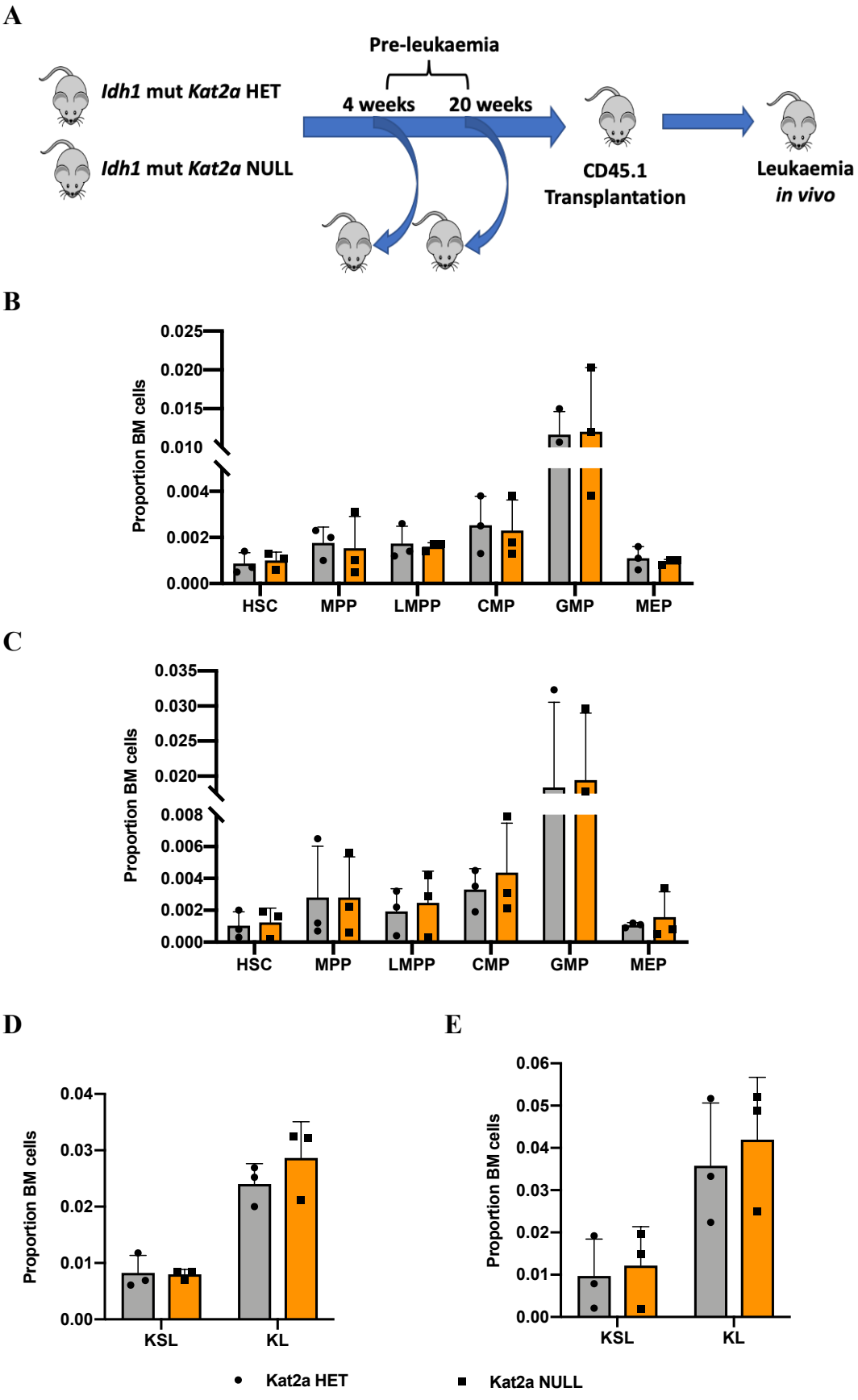

A
Pre-leukaemia
4 weeks
20 weeks
Idh1 mut Kat2a HET
Idh1 mut Kat2a NULL
CD45.1
Transplantation
Leukaemia
in vivo
B
Proportion BM cells
0.025
0.020
0.015
0.010
0.004
0.002
0.000
HSC
MPP
LMPP
CMP
GMP
MEP
C
Proportion BM cells
0.035
0.030
0.025
0.020
0.008
0.006
0.004
0.002
0.000
HSC
MPP
LMPP
CMP
GMP
MEP
D
Proportion BM cells
0.04
0.03
0.02
0.01
0.00
KSL
KL
E
Proportion BM cells
0.06
0.05
0.04
0.03
0.02
0.01
0.00
KSL
KL
Kat2a HET
Kat2a NULL

Figure 3.5: Functional characterization of *Idh1*R132H model.

(A) Schematic diagram representing experimental set-up. Bone marrow cells from *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals post 20 weeks of *pIpC* treatment were transplanted into lethally irradiated CD45.1 animals (n=8 animals/genotype). Pre-leukaemia studies were done post 4 and 20 weeks of weeks of *pIpC* treatment (n=3/genotype at each time point), Flow cytometry analysis for bone marrow cells obtained from *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals at **(B)** 4 weeks post *pIpC* (n=3/ genotype, mean ± SD), **(C)** 20 weeks post *pIpC* (n=3/genotype, mean ± SD) representing different compartments of haematopoietic hierarchy, Separate plot representing different lineage negative populations from the same flow cytometry analysis post **(D)** 4 weeks of *pIpC* (n=3/genotype, mean ± SD), **(E)** 20 weeks of *pIpC* (n=3/genotype, mean ± SD). (HSC- $Lin^-Kit^+Sca1^+CD34^-CD135^-$, MPP- $Lin^-Kit^+Sca1^+CD34^+CD135^-$, LMPP- $Lin^-Kit^+Sca1^+CD34^+CD135^+$, CMP- $Lin^-Kit^+Sca1^-CD34^{+/low}CD16/32^{low}$, GMP- $Lin^-Kit^+Sca1^+CD34^+CD16/32^{high}$, MEP- $Lin^-Kit^+Sca1^+CD34^-CD16/32^-$, KSL- $Lin^-Kit^+Sca1^+$, KL- Lin^-Kit^+).

The flow cytometry pre-leukaemia analysis for *Idh1*mut transformed cells was complemented with measurement of spleen and liver weights in order to assess early pre-leukaemia infiltration events (Fig 3.6). At 4 weeks post *pIpC* (n=3 for *Kat2a* HET and n=2 for *Kat2a* NULL), no differences in spleen weights were observed upon loss of *Kat2a* indicating no differences in pre-leukaemia burden (Fig 3.6A). This finding was recapitulated at 20 weeks post *pIpC* (n=3/genotype), where again no differences were observed in spleen weight upon *Kat2a* loss (Fig 3.6A). However, combining the analyses from 4 weeks and 20 weeks post *pIpC,* there was an increasing trend in spleen weight from 4 weeks to 20 weeks irrespective of genotype (Fig 3.6A), compatible with the inferences from Sasaki and colleagues, where a gain in spleen weight was observed with pre-leukaemia progression (Sasaki *et al.*, 2012b). Similarly, 4 weeks post *pIpC* (n=2/genotype), no changes were observed in liver weight upon loss of *Kat2a* (Fig 3.6B), similar to the findings at 20 weeks post *pIpC* (n=3/genotype) (Fig 3.6B). In contrast to the findings for spleen weight, no such trend was observed for alteration in liver weight with pre-leukaemia progression (Fig 3.6B).

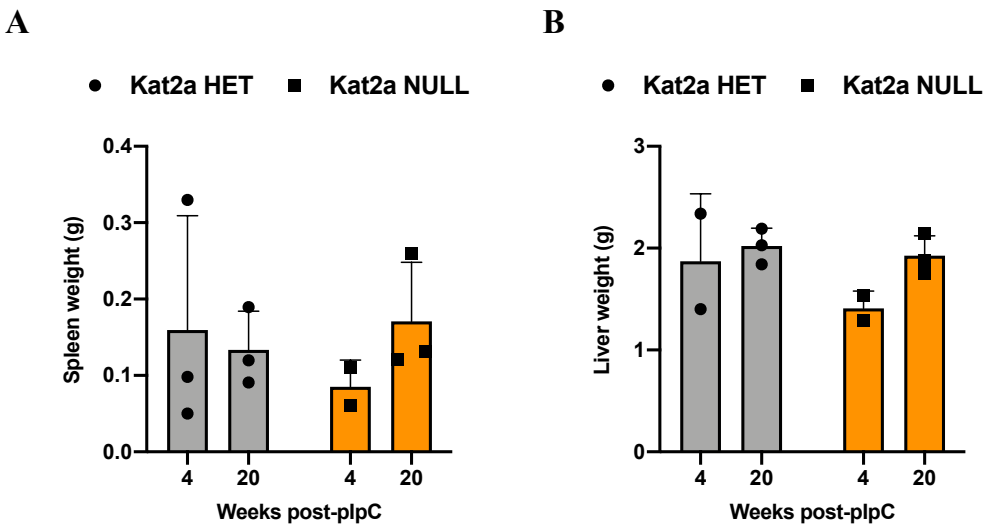


Figure 3.6: Spleen and liver weight during *Idh1*R132H pre-leukaemia.

Spleen weight at **(A)** 4 and 20 weeks (n=3 for *Kat2a* HET at both time points, n=2 for *Kat2a* NULL at 4 weeks, mean ± SD), **(B)** 4 and 20 weeks (n=3/genotype at each time point, mean ± SD)**,** post *pIpC.*

3.5 Loss of *Kat2a* aids in *Idh1*R132H transformation during pre-leukaemia

As loss of *Kat2a* aids in self-renewal of *RUNX1-RUNX1T1(9a)* pre-leukaemia, we wanted to study whether this observation can be recapitulated in *Idh1*mut model. For this, BM cells obtained from *Kat2a* HET and *Kat2a* NULL animals with *Idh1*mut at 4 weeks and 20 weeks post *pIpC* treatment were maintained in a colony forming assay at 50,000 cells/condition in technical duplicates (n=3/genotype for each time point). The colonies were scored every 7-10 days and re-plated serially at 10,000 cells/condition. There was an increase in re-plating efficiency in *Kat2a* NULL cells post 4 weeks *pIpC* suggesting that *Kat2a* plays a general role in promoting self-renewal of transformed cells at early pre-leukaemia stage (Fig 3.7A). On the other hand, post 20 weeks of *pIpC*, the difference in self-renewal potential between *Kat2a* NULL and *Kat2a* HET was lost suggesting that it was an early effect of the loss of *Kat2a* (Fig 3.7B).

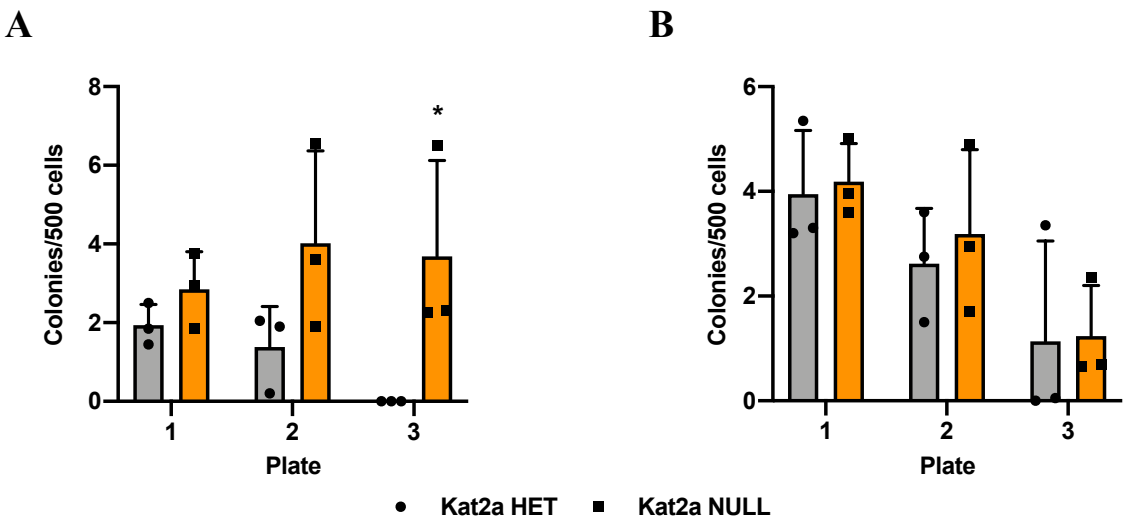


Figure 3.7: **Colony forming unit analysis for *Idh1*R132H pre-leukaemia.**

(A) Serial re-plating of colony forming cell assay from pre-leukaemia bone marrow cells obtained from *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals post 4 weeks of *pIpC* treatment (n=3 animals/genotype, mean ± SD, Student's t-test p=0.037*) **(B)** Serial re-plating of colony forming cell assay from pre-leukaemia bone marrow cells obtained from *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals post 20 weeks of *pIpC* treatment (n=3 animals/genotype, mean ± SD)

3.6 *Idh1*R132H animals showed myeloproliferation but did not develop leukaemia

To understand whether the gain in self-renewal at early stages of pre-leukaemia transformation impacts overall leukaemia progression, I started with *Kat2a* HET and *Kat2a* NULL bone marrow cells (n=3/genotype) studied at 4 weeks post *pIpC*. $1*10^6$ cells from each of the *Kat2a* HET and *Kat2a* NULL animals were injected into each lethally irradiated CD45.1 recipient such that there were 3 representative recipients from each donor. Unfortunately, most of the animals lost >50% of weight after -experimental set up and the experiment had to be terminated as per the study plan regulations. Further, an attempt was made by making use of *Kat2a* HET and *Kat2a* NULL BM cells (n=3/genotype) studied post 20 weeks of *pIpC*. The transplants were set up in a similar manner with recipients being sub-lethally irradiated.

To study differences in various haematological parameters upon loss of *Kat2a* during disease progression, peripheral blood sampling was done on a regular basis (Fig 3.8). There was an

increasing trend in WBC 38 weeks post transplantation in *Idh1*R132H *Kat2a* NULL animals compared to *Idh1*R132H *Kat2a* HET animals, perhaps highlighting an acceleration in disease progression (Fig 3.8A). However, the number of WBCs were found to be lower at 51 weeks post transplantation in both the genotypes which may indicate a technical bias. In contrast to this, the HGB levels remained stable post transplantation in both the genotypes with a slight increase at 10 weeks post transplantation in both the genotypes (Fig 3.8B). The levels of platelets in the peripheral blood were mostly fluctuating with a slight increase at 10-, 15- and 38-weeks post transplantation. This could be a result of technical issues while analysing the samples. Overall, no differences were observed in platelets levels in both genotypes (Fig 3.8C). Altogether, the peripheral blood analysis suggested that loss of *Kat2a* did not lead to any changes in the blood composition during the progression of *Idh1*R132H pre-leukaemia. This finding was in line with the insights obtained from *RUNX1-RUNX1T1(9a)* model.

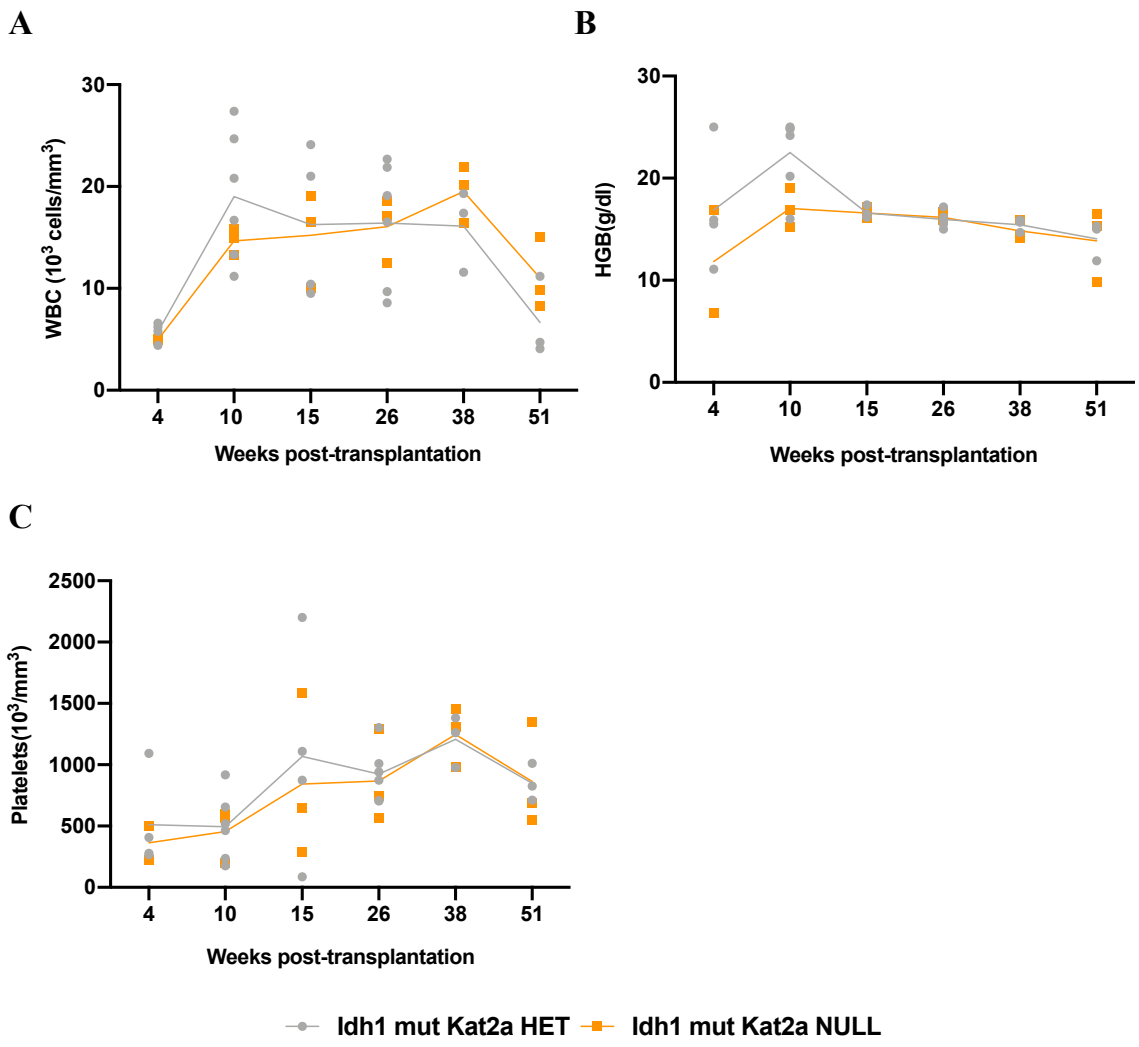


Figure 3.8: Peripheral blood analysis.

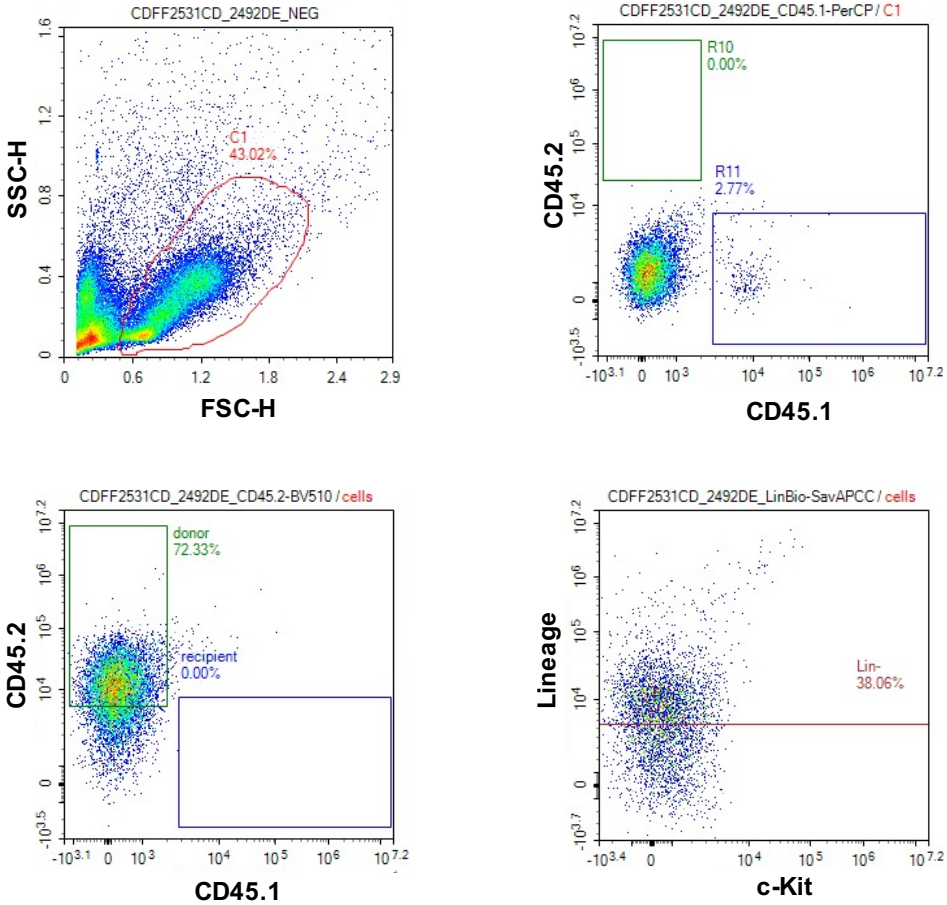
Peripheral blood analysis to study different haematological parameters, including, **(A)** White Blood Cells, **(B)** Haemoglobin, **(C)** Platelets for both *Idh1*mut *Kat2a* HET and *Idh1*mut *Kat2a* NULL animals (plots represent analysis for n=6 animals for *Idh1*mut *Kat2a* HET and n=3 animals for *Idh1*mut *Kat2a* NULL at 10 weeks post-transplantation, mean ± SD).

After transplantation, the animals were routinely observed for the presence of clinical symptoms of hunched posture, inappetence and lethargy, indicative of leukaemia development. However, none of the animals developed such symptoms until 52 weeks post transplantation. The animals which died of reasons other than leukaemia during this period were excluded from further analysis. The animals which were alive until 52 weeks post transplantation were culled using schedule 1 method of euthanasia; BM cells and spleen cells were isolated and stained with a cocktail of antibodies and flow cytometry analysis was performed (Methods)

(n=3/genotype). These antibodies include CD45.1 (recipient animals), CD45.2 (donor animals), lineage cocktail (Mac1, Gr1, CD3e, Ter119, B220) to exclude terminally differentiated cells, Sca1, c-Kit, to identify stem and progenitor compartment of cells, FcγR (myeloid marker) and Cd34 to identify HSCs and Flt3 which is short term HSCs/multipotent progenitor marker. The experiment was performed by Caitlin Cash, a master's student in Pina group.

The flow cytometry analysis suggested that majority of cells present in the bone marrow were CD45.2 positive rather than being CD45.1 positive, highlighting successful transplantation of *Idh1*R132H transformed cells in both *Kat2a* HET and *Kat2a* NULL animals (Fig 3.9A.I, B.I). All the animals transplanted with *Idh1*R132H transformed cells either with *Kat2a* HET or *Kat2a* NULL showed more than 50% of the bone marrow cells which were CD45.2 positive. We then focussed on CD45.2 positive cells for further analysis. In order to understand whether these cells have enrichment for terminally differentiated cells, we looked at lineage markers which suggested nearly 40-50% of the cells were lineage negative in case of both *Kat2a* HET or *Kat2a* NULL animals (Fig 3.9A.II, B.II) and thus indicate presence of early progenitor cells in the bone marrow. We then further narrowed down our analysis to look at the presence of c-Kit and Sca-1 marker expression specifically within the lineage negative compartment of cells in order to distinguish between progenitor and stem cells. We observed the presence of progenitor cells characterized by $Lin^{-}c\text{-}Kit^{+}Sca1^{-}$ (KL) which are further capable of expansion, compared to $Lin^{-}c\text{-}Kit^{+}Sca1^{+}$ (KLS) population representing stem compartment of cells (Fig 3.9A.III, B.III). The KL population of cells were further segregated on the basis of the presence of FcγR and Cd34 expression in these cells. The presence of $Fc\gamma R^{+}Cd34^{+}$ population of cells would have indicated GMP-like population, however, Cd34 expression was not observed in these cells, compatible with the absence of leukaemia phenotype. This was consistent in both *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL cells. However, both genotypes showed an enrichment in $Fc\gamma R^{+}$ population of cells suggesting an accumulation in myeloid progenitors and indicative of myeloproliferation (Fig 3.9A.IV, B.IV).

A


CDFF2531CD_2492DE_NEG
SSC-H
FSC-H
C1
43.02%
CDFF2531CD_2492DE_CD45.1-PerCP / C1
CD45.2
CD45.1
R10
0.00%
R11
2.77%
CDFF2531CD_2492DE_CD45.2-BV510 / cells
CD45.2
CD45.1
donor
72.33%
recipient
0.00%
CDFF2531CD_2492DE_LinBio-SavAPCC / cells
Lineage
c-Kit
Lin-
38.06%

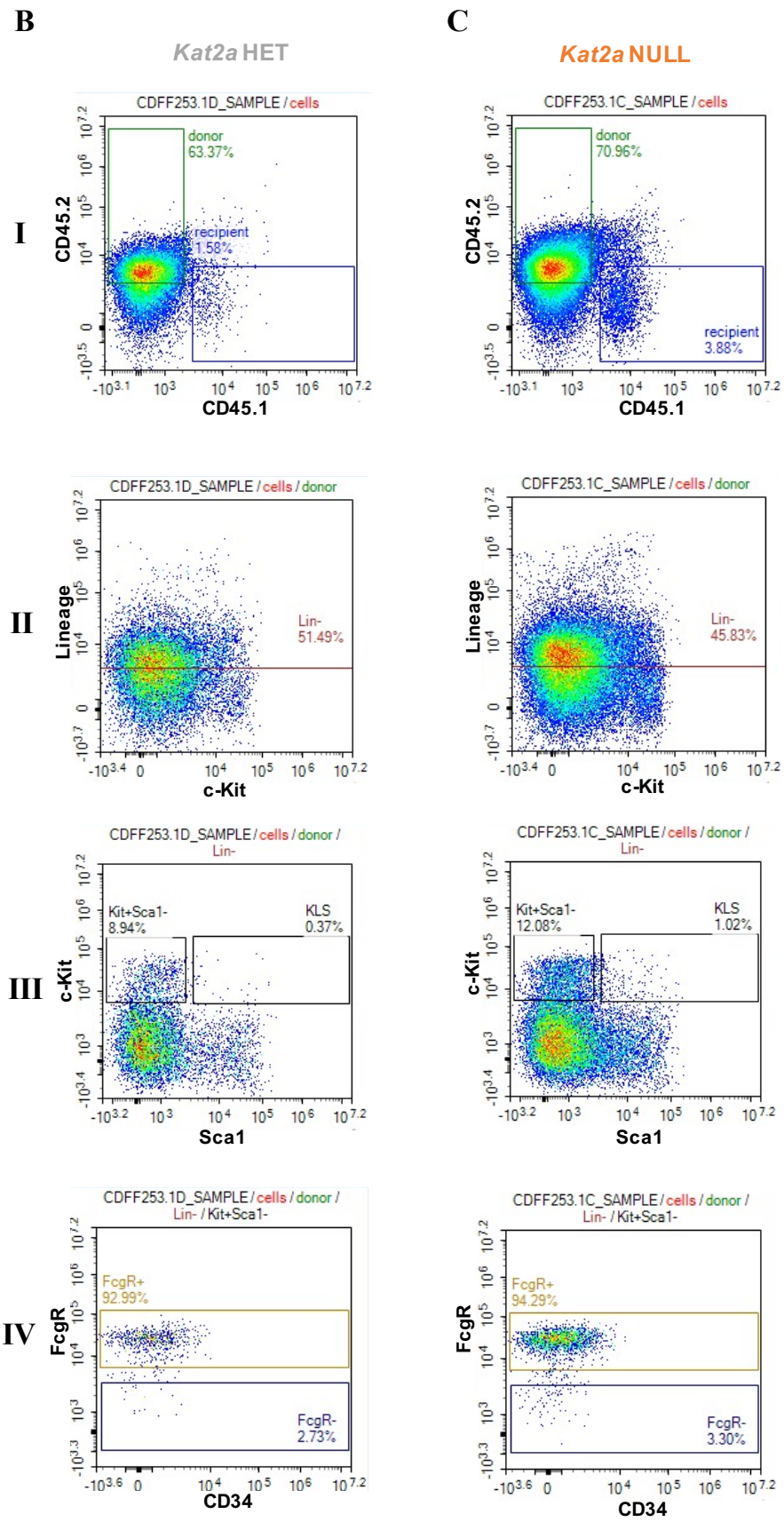

B
C
Kat2a HET
Kat2a NULL
I
II
III
IV
CDFF253.1D_SAMPLE / cells
donor 63.37%
recipient 1.58%
CD45.2
CD45.1
CDFF253.1C_SAMPLE / cells
donor 70.96%
recipient 3.88%
CDFF253.1D_SAMPLE / cells / donor
Lin- 51.49%
Lineage
c-Kit
CDFF253.1C_SAMPLE / cells / donor
Lin- 45.83%
CDFF253.1D_SAMPLE / cells / donor / Lin-
Kit+Sca1- 8.94%
KLS 0.37%
Sca1
CDFF253.1C_SAMPLE / cells / donor / Lin-
Kit+Sca1- 12.08%
KLS 1.02%
CDFF253.1D_SAMPLE / cells / donor / Lin- / Kit+Sca1-
FcgR+ 92.99%
FcgR- 2.73%
FcgR
CD34
CDFF253.1C_SAMPLE / cells / donor / Lin- / Kit+Sca1-
FcgR+ 94.29%
FcgR- 3.30%

Figure 3.9: Flow cytometry analysis of *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL animals.

(A) Flow cytometry schematic control plots for *Idh1R132H* transplants**, (B)** Flow cytometry plots for *Idh1*R132H transformed *Kat2a* HET cells 52 weeks post transplantation highlighting plots between CD45.1 and CD45.2 (I), c-Kit and Lineage (II), Sca-1 and c-Kit (III) and Cd34 and FcgR (IV), **(C)** Flow cytometry plots for *Idh1*R132H transformed *Kat2a* NULL cells 52 weeks post transplantation highlighting plots between CD45.1 and CD45.2 (I), c-Kit and Lineage (II), Sca-1 and c-Kit (III) and Cd34 and FcgR (IV) (KLS- $Lin^{-}c\text{-}Kit^{+}Sca1^{+}$, Lin- Lineage), (Plots mentioned in II, III and IV represent percentage population of cells within CD45.2 population).

Overall, the phenotypic characterization of cells did not reveal any significant differences upon loss of *Kat2a* along with any changes in disease progression suggesting that loss of *Kat2a* is not sufficient to carry out leukaemia development. Nonetheless, the analysis confirmed that *Idh1* mutation represents pre-leukaemia phase which requires the presence of collaborating mutations for leukaemia development.

3.7 *Kat2a* loss promotes enrichment of c-$Kit^{+}Mac1^{-}$ progenitor cells in *Idh1*R132H transplants

After characterizing the phenotype of *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL transplanted cells using flow cytometry, we then studied the self-renewal potential of these cells in order to understand whether *Kat2a* NULL cells show an advantage in their replating capacity. For this, both *Kat2a* HET and *Kat2a* NULL cells transformed with *Idh1*R132H were plated into methylcellulose based semi-solid medium at 10,000 cells/condition (n=3 animals/genotype) and colonies were scored after 7-10 days. As mentioned earlier, the present model has a 30% leukaemia incidence 1-year post-mutation activation, so the bone marrow cells obtained from *Idh1*R132H leukaemia animals from collaborators at Sanger institute served as control. At plating one, there was an increasing trend in colony forming potential in *Idh1*R132H transformed *Kat2a* NULL cells compared to *Kat2a* HET. On the other hand, leukaemia cells had highly variable colony forming potential in different replicates (Fig 3.10A). However, at plating 2 there was an increase in colony forming potential upon *Kat2a* loss in *Idh1*R132H transformed cells compared to *Kat2a* HET (Fig 3.10A) suggesting that loss of *Kat2a* provides a self-renewal potential to these transplanted cells. Moreover, the self-

renewal potential observed in *Kat2a* NULL cells transformed with *Idh1*R132H was comparable to the colony forming potential of *Idh1*R132H leukaemia cells, altogether suggesting that although we did not capture leukaemia progression in our transplants, loss of *Kat2a* provides a self-renewal advantage to *Idh1*R132H transformed cells which may consequentially lead to an accelerated leukaemic progression.

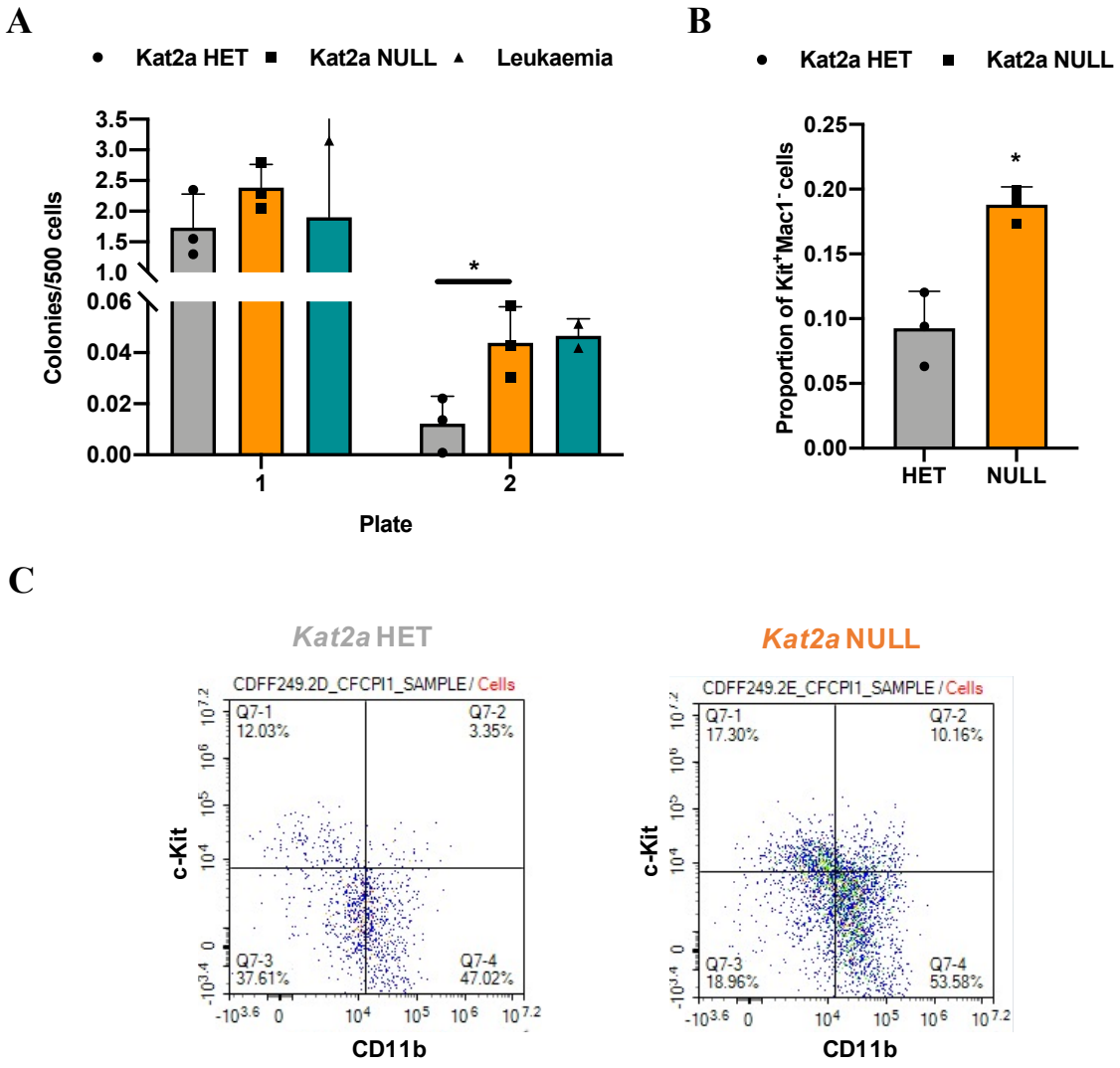


Figure 3.10: Functional characterization of *Idh1*R132H transplants.

(A) Serial re-plating of colony forming cell assay using bone marrow cells obtained from *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL animal transplants (n=3/genotype) along with *Idh1*R132H leukaemia animals (n=2) (mean ± SD, Student's t-test p=0.0398* for *Kat2a* HET vs *Kat2a* NULL), **(B)** Proportion of c-Kit^{+} Mac1^{-} cells in cells obtained from *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL bone marrow at plating 1 (mean ± SD, Student's t-test p=0.0152*), **(C)** Representative flow

cytometry plots highlighting c-Kit and Mac1 cell population in *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL bone marrow cells at plating 1.

To characterize the progenitor population of cells contributing to the colony forming potential, we performed flow cytometry analysis from the cells obtained after plating one. The flow cytometry analysis was gated on cells which were characterized as donor cells based on CD45.2 expression. The cells were stained with all the lineage markers as well as c-Kit marker in order to assess presence of a progenitor phenotype. Overall, no differences were observed in any of the haematopoietic compartments upon *Kat2a* loss, however, a combined analysis of proportion of c-Kit^{+}Mac1^{-} population with the colony forming cells obtained after plating one suggested an increase in c-Kit^{+}Mac1^{-} population of cells upon loss of *Kat2a* (Fig 3.10B). The representative plots for c-Kit^{+}Mac1^{-} population in *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL cells are shown in Fig 3.10C.

3.8 Loss of *Kat2a* does not impact DNA damage in *Idh1*R132H pre-leukaemia

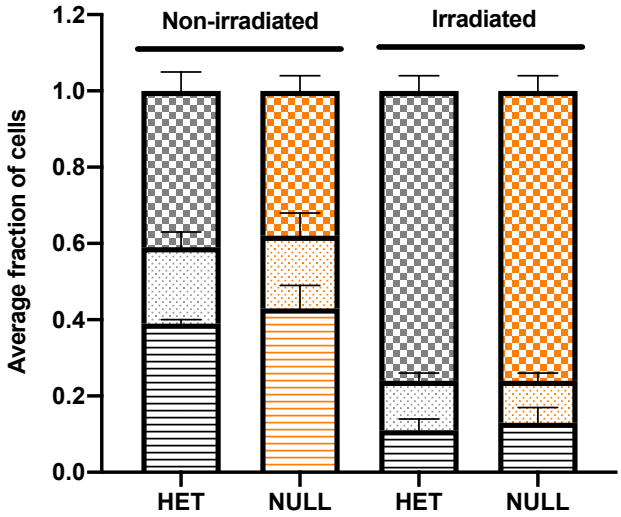
Combined observations from *RUNX1-RUNX1T1(9a)* and *Idh1*R132H pre-leukaemia models suggested that loss of *Kat2a* at early stages of disease progression aids in transformation by increasing self-renewal capacity of these cells. This enhanced self-renewal capacity is overall advantageous in accelerating pre-leukaemia to leukaemia progression. Since these models are dependent on the acquisition of incorporating mutations, it was important to study whether loss of *Kat2a* is accelerating disease progression by increasing genomic instability.

For this, DNA damage analysis was conducted in the lab on *Idh1*mut transformed BM cells with *Kat2a* HET or *Kat2a* NULL genotype collected at 4 weeks and 20 weeks post *pIpC* treatment (n=3/genotype at each time point). For this, the collected BM cells were thawed and irradiated to induce DNA damage. Non-irradiated cells served as control. Cells were recovered post irradiation and subsequently stained with anti-phospho-Histone H2AX (γH2AX) along with Alexa Fluor 594 rabbit anti-mouse as a secondary control. γH2AX is formed upon phosphorylation of Ser140 as a functional consequence of the DNA damage response and act as a surrogate for DNA double stranded breaks (Horton, 2017). DAPI was used as a nuclei stain and acquisition was done on a confocal microscope (Methods). Cells were categorized

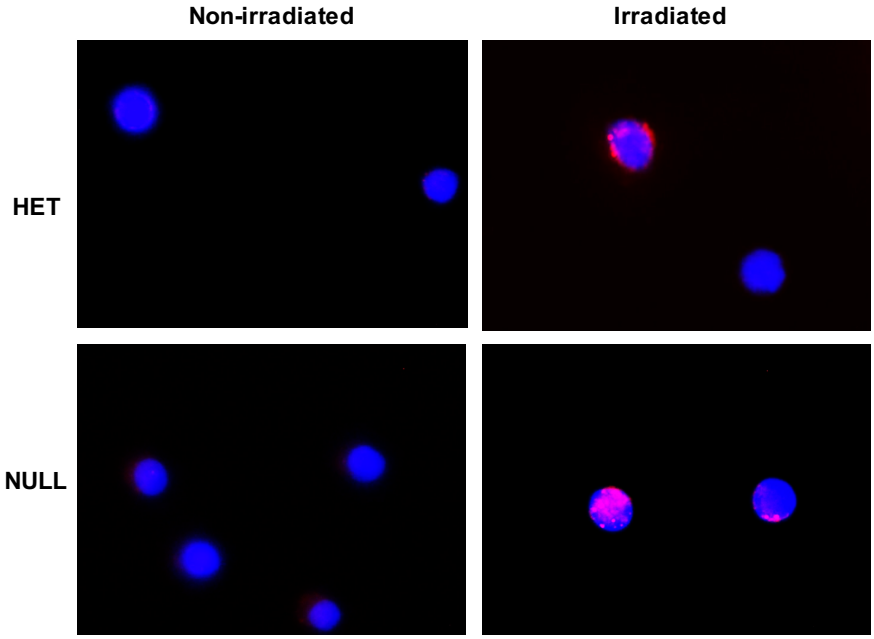
into ones with 0 foci, 1-5 foci, or more than 5 foci. Cells with 0 foci represent ones with no DNA damage observed, 1-5 foci represent cells with moderate DNA damage, and more than 5 foci represent cells with high DNA damage. At least 100 cells were scored for each sample from three individual replicates.

Upon irradiation, there was an increase in the cell population with more than 5 foci compared to the non-irradiated cell population confirming the DNA damage induced by irradiation (Fig 3.11A and 3.11C). However, there were no significant differences in the number of cells showing more than 0 foci in *Kat2a* NULL as compared to *Kat2a* HET at both 4 weeks (Fig 3.11A and 3.11B) and 20 weeks post *pIpC* (Fig 3.11C and 3.11D). This suggested the potential involvement of non-genetic factors in accelerating leukaemia progression upon loss of *Kat2a*.

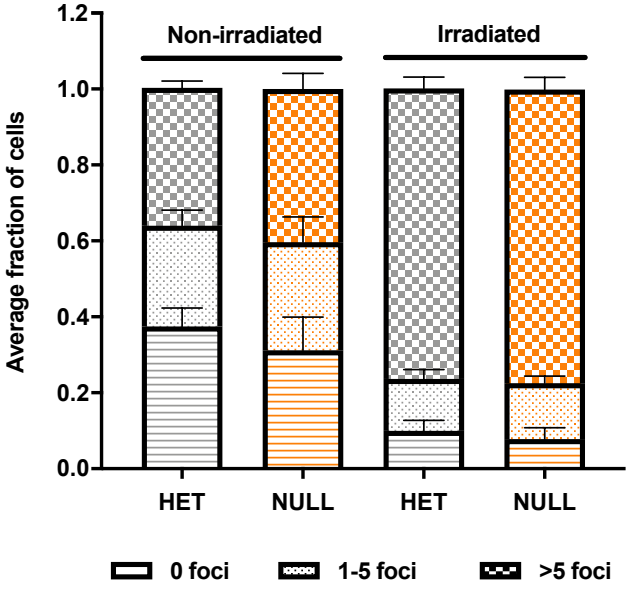
A


Non-irradiated
Irradiated
Average fraction of cells
1.2
1.0
0.8
0.6
0.4
0.2
0.0
HET
NULL
HET
NULL

B


Non-irradiated
Irradiated
HET
NULL

C


Non-irradiated
Irradiated
Average fraction of cells
1.2
1.0
0.8
0.6
0.4
0.2
0.0
HET
NULL
HET
NULL
0 foci
1-5 foci
>5 foci

D

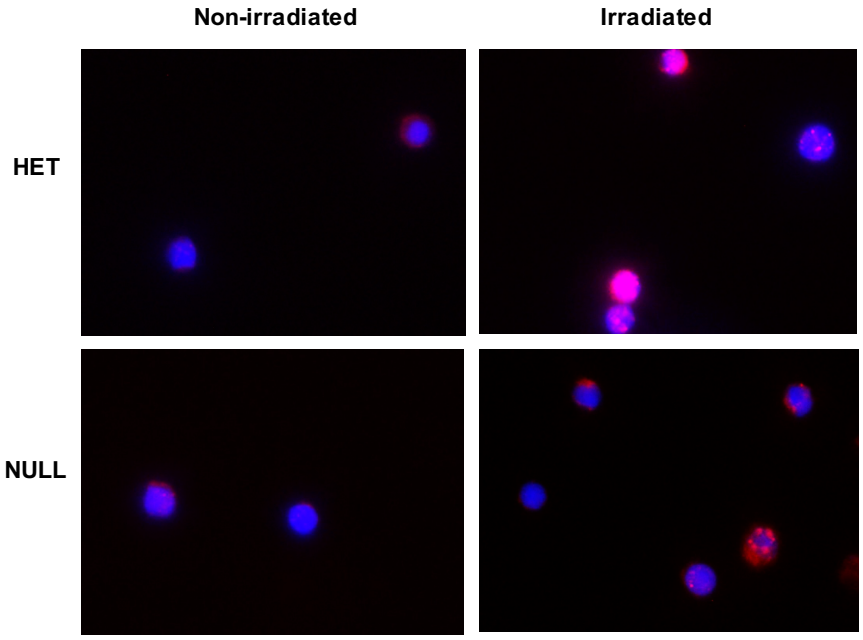

Non-irradiated
Irradiated
HET
NULL

Figure 3.11: DNA damage assay for *Idh1*R132H pre-leukaemia.
(A) Plot representing proportion of cells with γH2AX foci upon irradiation with 0 foci/cell, 1-5 foci/cell and >5 foci/cell in bone marrow cells obtained from *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals post 4 weeks of *pIpC* treatment where non-irradiated cells served as control (n= 3 animals/genotype, mean ± SD), **(B)** Representative microscopic images of γH2AX foci in *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals post 4 weeks of *pIpC* treatment, **(C)** Plot representing proportion of cells with γH2AX foci upon irradiation with 0 foci/cell, 1-5 foci/cell and >5 foci/cell in bone marrow cells obtained from *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals post 20 weeks of *pIpC* treatment (n=3 animals/genotype, mean ± SD), **(D)** Representative microscopic images of γH2AX foci in *Idh1*R132H *Kat2a* HET and *Idh1*R132H *Kat2a* NULL animals post 20 weeks of *pIpC* treatment. (*Idh1*R132H *Kat2a* HET- HET, *Idh1*R132H *Kat2a* NULL- NULL)

In this chapter, I have studied the functional consequences of loss of *Kat2a* during *RUNX1-RUNX1T1(9a)* and *Idh1*R132H pre-leukaemia progression. The loss of *Kat2a* was found to promote the *RUNX1-RUNX1T1(9a)* leukaemia progression with a significant number of cells being c-Kit$^+$, suggesting the presence of early progenitor cells prone to leukaemic transformation. The leukaemia analysis was performed in two separate cohorts of animals with varying transduction efficiency, however, in both cases a consistent observation of accelerated leukaemia progression was observed upon *Kat2a* loss. Phenotypic characterization of individual leukaemia animals highlighted a different proportion of leukaemia markers including presence of stem (c-Kit$^+$Sca1$^+$) and progenitor cells (c-Kit$^+$Sca1$^-$), confirming the multi-hit requirement for *RUNX1-RUNX1T1(9a)* leukaemia progression. The infiltration of leukaemia cells was observed in spleen as well as in liver in some cases. Strikingly, terminal blood analysis did not reflect any differences in leukaemia burden upon loss of *Kat2a*, compatible with our previous lab observation in *MLL-AF9* leukaemia (Domingues *et al.*, 2020).

Upon studying *RUNX1-RUNX1T1(9a)* pre-leukaemia stages, a perpetuation in *RUNX1-RUNX1T1(9a)*$^+$ pre-leukaemia clones was observed in the peripheral blood 12 weeks and 17 weeks post-transplantation. Due to a smaller number of animals left post 17 weeks of transplantation, the analysis could not be performed with confidence, however, an increasing trend in GFP$^+$ population was observed upon loss of *Kat2a.* In line with this, the flow cytometry characterization of pre-leukaemia animals 2 months and 4 months post transplantation

indicated an increase in $GFP^+c\text{-}Kit^+Fc\gamma R^+$ cell population upon *Kat2a* loss, indicative of accumulation of myeloid progenitor cells. The accumulation of myeloid progenitors was compatible with the gradual progression of *RUNX1-RUNX1T1(9a)* pre-leukaemia. However, the haematological parameters obtained from peripheral blood sampling did not show any difference upon loss of *Kat2a,* suggesting that loss of *Kat2a* does not have any impact on haematological parameters globally. However, the variability observed at different time points may be suggestive of technical bias during sampling, given that it follows similar trend in both the genotypes.

The pre-leukaemia cells obtained from both *Kat2a* WT and *Kat2a* NULL cells 2 months and 4 months post transplantation were assessed for their self-renewal potential using colony forming assay. The analysis suggested an increase in colony forming potential associated with *Kat2a* NULL cells compared to *Kat2a* WT cells with a trend at 2 months and at 4 months post transplantation. The variability in colony forming efficiency observed at 2 months was due to the varying GFP levels in individual samples. Similar analysis was performed using spleen cells obtained from *Kat2a* WT and *Kat2a* NULL animals 2 months post transplantation and an increase in colony forming potential was observed in *Kat2a* NULL cells, compatible with the observations from bone marrow cells (data not shown). Similar insights were obtained from colony forming assay performed on *Kat2a* WT and *Kat2a* NULL cells transformed with *RUNX1-RUNX1T1(9a) in vitro.* Interestingly, the pro-leukaemogenic effect of *Kat2a* loss was specific to disease initiation as the ablation of *Kat2a* in established *RUNX1-RUNX1T1(9a)* AML cells resulted in reduced maintenance of transformed cells, in line with my lab's observations in AML cell lines (Tzelepis *et al.*, 2016) and the mouse *MLL-AF9* AML (Domingues *et al.*, 2020).

In order to study whether the functional consequences associated with *Kat2a* loss observed in *RUNX1-RUNX1T1(9a)* pre-leukaemia are model specific, another model of pre-leukaemia, namely *Idh1*R132H was studied in a *Kat2a* conditional knockout background. The *Idh1*R132H pre-leukaemia analysis was performed 4 weeks and 20 weeks post initiation of mutation recombination event as confirmed by PCR (Methods). Phenotypic characterization of pre-leukaemia bone marrow cells revealed no significant differences upon loss of *Kat2a* at both 4-week and 20-week time points, suggesting that loss of *Kat2a* does not impact overall bone

marrow cellularity during *Idh1*R132H pre-leukaemia, an observation in line with observations in models previously studied in the lab, which include *RUNX1-RUNX1T1(9a)* as well as *MLL-AF9* model of leukaemia. There were no significant changes in proportion of HSCs upon *Kat2a* loss in *Idh1*R132H pre-leukaemia at both time points. However, between 4 weeks to 20 weeks, an increasing trend in overall cellularity was observed, in particular for the CMP and GMP population irrespective of genotype. Similarly, no significant differences were found between the genotypes at 4 weeks and 20 weeks in KSL and KL populations, suggesting that loss of *Kat2a* does not impact stem and progenitor compartment during *Idh1*R132H pre-leukaemia progression. However, while comparing these populations from 4 weeks to 20 weeks, there was an increasing trend in KL at 20 weeks indicating putative pre-leukaemia progression. This observation was compatible with the increase in $GFP^{+}c\text{-}Kit^{+}Fc\gamma R^{+}$ from 2 months to 4 months during *RUNX1-RUNX1T1(9a)* pre-leukaemia progression. Although no infiltration of pre-leukaemia cells was observed in spleen and liver, however, an increasing trend in spleen weight from 4 weeks to 20 weeks was observed irrespective of genotype, potentially indicative of pre-leukaemia burden.

The pre-leukaemia cells obtained from *Kat2a* WT and *Kat2a* NULL cells 4 weeks and 20 weeks post transplantation were assessed for their self-renewal potential using colony forming assay. An increase in colony forming efficiency was observed upon *Kat2a* loss 4 weeks post initiation of mutation recombination event. This gain in self-renewal potential was not maintained until 20 weeks. This analysis could have benefitted from including pre-leukaemia analysis at a time point between 4 weeks and 20 weeks of pre-leukaemia progression to understand the hierarchical progression of gain in self-renewal upon *Kat2a* loss during *Idh1*R132H pre-leukaemia progression. Nonetheless, the analysis altogether suggested that loss of *Kat2a* aids a transient increase in colony forming potential during early stages of pre-leukaemia.

In order to understand whether the gain in self-renewal at early stages of pre-leukaemia transformation impacts overall leukaemia progression, CD45.1 animals were transplanted with *Idh1*R132H transformed *Kat2a* HET and *Kat2a* NULL bone marrow cells. An increase in WBCs was observed at 10 weeks in both the genotypes with an increasing trend specifically in *Kat2a* NULL cells at 38 weeks, perhaps indicating the accelerated establishment of leukaemia

upon loss of *Kat2a*. In contrast, the HGB levels and platelets remained stable post transplantation in both genotypes. Altogether, the peripheral blood analysis suggested that loss of *Kat2a* did not lead to any changes in the blood composition during the progression of *Idh1*R132H pre-leukaemia. This finding was in line with insights obtained from *RUNX1-RUNX1T1(9a)* model. Flow cytometry based phenotypic characterization of *Idh1*R132H *Kat2a* HET and *Kat2a* NULL animals did not show changes in any of the haematopoietic compartments suggesting that loss of *Kat2a* does not impact overall cellularity upon *Idh1*R132H mediated transformation which was compatible with pre-leukaemia analysis. None of the animals progressed to leukaemia but showed a development of myeloproliferation, suggesting the requirement of additional incorporating mutations for leukaemia progression, in line with the only model published so far (Sasaki *et al.*, 2012b). However, combining the serial replating assay with flow cytometry characterization, we observed an increase in c-$Kit^{+}Mac1^{-}$ population of cells upon *Kat2a* loss, suggesting that loss of *Kat2a* may promote the enrichment of progenitor cells which are likely capable of expansion during the process of leukaemia progression.

Since both *RUNX1-RUNX1T1(9a)* and *Idh1*R132H models of pre-leukaemia are described as multi-hit haematological malignancies which require incorporation of additional mutations in order for leukaemia establishment, we further studied whether loss of *Kat2a* accelerates disease progression by causing genomic instability. Upon irradiation, there was an increased population of cells having γH2AX foci indicating the induction of DNA damage upon irradiation. However, no significant differences were observed in *Kat2a* NULL cells transformed with *Idh1*R132H compared to *Kat2a* HET cells at both 4 weeks and 20 weeks suggesting that loss of *Kat2a* does not increase DNA damage by inducing double strand breaks. The analysis altogether indicated that the accelerated pre-leukaemia progression upon loss of *Kat2a* is likely due to the potential involvement of non-genetic factors. The next chapter discusses the single cell RNA sequencing (scRNA-seq) performed on *RUNX1-RUNX1T1(9a)* pre-leukaemia cells studied 2 months and 4 months post transplantation. The chapter highlights the scRNA-seq pre-processing, filtering, and analysis in order to study the transcriptional programmes impacted by loss of *Kat2a* during *RUNX1-RUNX1T1(9a)* pre-leukaemia.

4 Identification of transcriptional programmes associated with *Kat2a* loss in *RUNX1-RUNX1T1(9a)* pre-leukaemia

The previous chapter discussed the functional role of *Kat2a* in promoting pre-leukaemia transformation *in vivo* in both *RUNX1-RUNX1T1(9a)* as well as *Idh1R132H* models of AML. The *in vitro* analysis of these pre-leukaemia clones highlighted a perpetuation of these population of cells upon loss of *Kat2a*. These pre-leukaemia clones were further found to be associated with an enhanced self-renewal capacity suggesting these cells have the capability to maintain the transformants, which may have further expanded and consequently led to an accelerated leukaemia progression as evidenced in the *RUNX1-RUNX1T1(9a)* model. As discussed previously, in order to understand whether this accelerated leukaemia progression is also a consequence of an increased DNA damage, γH2AX staining was done in *Idh1R132H* transformed *Kat2a* HET and *Kat2a* NULL cells. The analysis did not provide evidence for an increase in DNA damage upon *Kat2a* loss in case of *Idh1R132H* model, suggesting that genomic instability may not be the underlying cause of this accelerated leukaemia progression. These observations further highlighted the importance of understanding the role of non-genetic mechanisms during the process of pre-leukaemia transformation. Since *Kat2a* loss is central to an increase in cell-to-cell transcriptional variability (Moris *et al.*, 2018b; Domingues *et al.*, 2020), I tested the hypothesis by studying the potential link between loss of *Kat2a*, the associated increase in transcriptional variability, and accelerated pre-leukaemia progression using single cell RNA sequencing (scRNA-seq) of *RUNX1-RUNX1T1(9a)* pre-leukaemia cells. This chapter describes the scRNA-seq analysis, in particular the pre-processing of raw data obtained from scRNA-seq and further filtering and normalization steps which were required to generate a workable gene-expression matrix which contained high quality cells representing gene expression levels beyond the level of technical noise. Further, this chapter also includes the various differential gene expression comparisons between genotype-specific samples as well as different time points within the same genotype in order to identify instability in transcriptional programmes induced upon loss of *Kat2a* during pre-leukaemia progression.

4.1 Single-cell RNA sequencing of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT and *Kat2a* NULL pre-leukaemia cells

In order to study a potential link between loss of *Kat2a*, the associated increase in transcriptional variability, and accelerated pre-leukaemia progression, scRNA-seq was performed on *Kat2a* WT and *Kat2a* NULL cells transformed with *RUNX1-RUNX1T1(9a),* collected 2 months and 4 months post transplantation. scRNA-seq was performed using 10X genomics technology with 3' chemistry (v2). The single cell 3' expression profiling can be divided into four steps-

4.1.1 Gel Bead-In-EMulsions (GEMs) generation and Barcoding

The single cell 3' expression profiling utilizes a microfluidic platform for digital gene expression profiling by making use of 10X GemCode technology. It does so by partitioning thousands of cells into nanolitre-scale Gel Bead-In-EMulsions (GEMs), where all generated cDNA share a common 10X barcode sequence. Libraries are generated and sequenced from the cDNA and the 10X barcodes are used to associate individual reads back to the individual partitions. To achieve single cell resolution, the cells are delivered at a limiting dilution, such that the majority (~90-99%) of generated GEMs contain no cells, while the remainder largely contain a single cell. Upon dissolution of the single cell 3' Gel Bead in a GEM, primers containing (i) an Illumina R1 sequence (read1 sequencing primer), (ii) a 16 nt 10X barcode, (iii) a 10 nt Unique Molecular Identifier (UMI), and (iv) a poly-dT primer sequence are released and mixed with cell lysate and master mix. Incubation of the GEMs then produces barcoded, full-length cDNA from poly-adenylated mRNA. After incubation, the GEMs are broken, and the pooled fractions are recovered.

4.1.2 Post GEM-RT clean up and cDNA amplification

Silane magnetic beads are used to remove leftover biochemical reagents and primers from the post GEM reaction mixture. Full length barcoded cDNA is then amplified by PCR to generate sufficient mass for library construction.

4.1.3 Library construction

Enzymatic fragmentation and size selection are used to optimize the cDNA amplicon size prior to library construction. R1 (read 1 primer sequence) are added to the molecules during GEM incubation. Primers used in Illumina bridge amplification (P5, P7), a sample index, and R2 (read 2 primer sequence) are added during library construction via end repair, A tailing, adaptor ligation, and PCR.

4.1.4 Sequencing libraries

The single cell 3' protocol produces Illumina-ready sequencing libraries. A single cell 3' library comprises standard Illumina paired-end constructs which begin and end with P5 and P7. The single cell 3' 16 bp 10X barcode and 10 bp UMI are encoded in Read 1, while Read 2 is used to sequence the cDNA fragment. Sample index sequences are incorporated as the i7 index read. Read 1 and Read 2 are standard Illumina sequencing primer sites used in paired-end sequencing.

The sequencing of a single cell 3' library produced a standard Illumina raw Base Call (BCL) data output folder. The BCL data included the paired-end Read 1 (containing the 16 bp 10x Barcode and 10 bp UMI) and Read 2, and the sample index in the i7 index read. In order to assess the quality of these BCL files, I conducted quality control using FastQC (Wingett and Andrews, 2018), filtering based on Phred score (Ewing and Green, 1998). As described in methods, a plot representing quality scores across individual bases was studied in order to get an overview of the range of quality values across all bases at each position in the FastQC file. The higher the score across the bases, the better the base call. In addition to this, the per sequence quality score report was also studied, which allowed identification of sequences having universally low-quality values. Post QC, the CellRanger (v2.2) pipeline developed by 10X genomics was followed in order to align reads and generate the gene-cell expression matrix. The process was initiated by running the cellranger mkfastq command which demultiplexes raw base call (BCL) files generated by Illumina sequencing into FASTQ files. These FASTQ files were further fed into the cellranger count pipeline which performs

alignment, filtering, barcode counting, and UMI counting. mm10 was used as a reference genome for alignment. These two pipelines were run individually for each sample in order to obtain good quality reads for further analysis.

In total, 1767 cell barcodes were subjected to sequencing: 379 cells from *Kat2a* WT at 2 months, 369 cells from *Kat2a* NULL at 2 months, 518 cells from *Kat2a* WT at 4 months, and 501 cells from *Kat2a* NULL at 4 months. Post QC and filtering, all 1767 cells were retained for downstream processing. Once the reads from these cell barcodes were aligned to the mm10 mouse reference genome using the cellranger count (v2.2) command (https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome), the cellranger aggr command was used to normalise all four samples to the same sequencing depth and recompute the gene-barcode matrices. The gene-barcode matrix generated had 1767 cells in total, 174770 mean reads per cell with 1575 median genes detected per cell. This resultant matrix was then used for further processing and analysis.

4.2 Single-cell RNA sequencing- pre-processing, filtering and normalization

Although cellranger differentiated between cells and technical noise and helped in removing background noise and dropout events from the gene expression matrix, certain biological parameters still needed to be imposed. For example, filtering parameters like the minimum number of cells expressing a gene or the minimum number of genes expressed in a cell could be varied to ensure stringent filtering for downstream analysis.

To ensure the gene-barcode matrix retains good quality cells and gene expression values, I imposed two different types of filtering-

1. First, I set the minimum number of genes expressed in a cell to 500 based on a previous lab study (Domingues *et al.*, 2020), where each cell that expressed less than 500 genes was considered a poor quality cell.

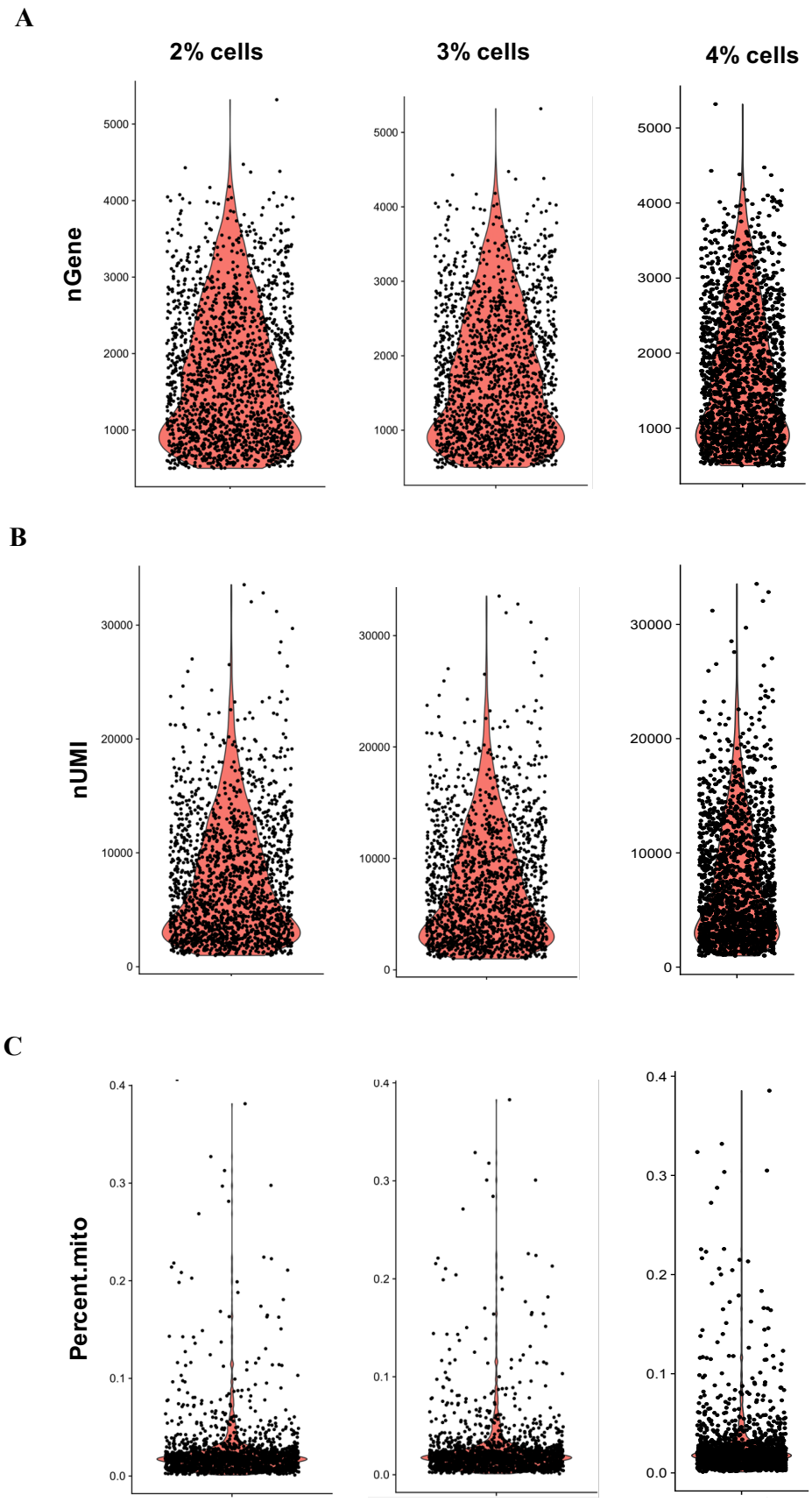

A
2% cells
3% cells
4% cells
nGene
5000
4000
3000
2000
1000
B
nUMI
30000
20000
10000
0
C
Percent.mito
0.4
0.3
0.2
0.1
0.0

Figure 4.1: Visualization of gene and cell counts using Seurat v2.4.
(A) Violin plot representing distribution of number of genes in cells in the dataset after imposing a cut-off of each gene being expressed in at least 2% of cells (left), 3% of cells (middle), and 4% of cells (right), and each cell having at least 500 genes, where each data point represents an individual cell, **(B)** Violin plot representing distribution of number of UMI molecules in the dataset after imposing a cut-off of each gene being expressed in at least 2% of cells (left), 3% of cells (middle) and 4% of cells (right), and each cell having at least 500 genes, where each data point represents an individual cell **(C)** Violin plot representing percentage of mitochondrial genes after imposing a cut-off of each gene being expressed in at least 2% of cells (left), 3% of cells (middle) and 4% of cells (right), and each cell having at least 500 genes, where each data point represents an individual cell.

2. Second, to remove technical noise in gene expression values, I imposed a cut-off where each gene is considered to have a significant expression in the given biological system only if a minimum number of cells express it. Since this is a critical step for obtaining rare cell populations which may have distinct gene expression signatures during pre-leukaemia progression, I started with 3 such cut-offs of 2% (36 cells), 3% (53 cells), and 4% (71 cells) for downstream analysis and decided the ideal cut-off for capturing such cell populations based on cell visualization and effective separation using dimensionality reduction analysis.

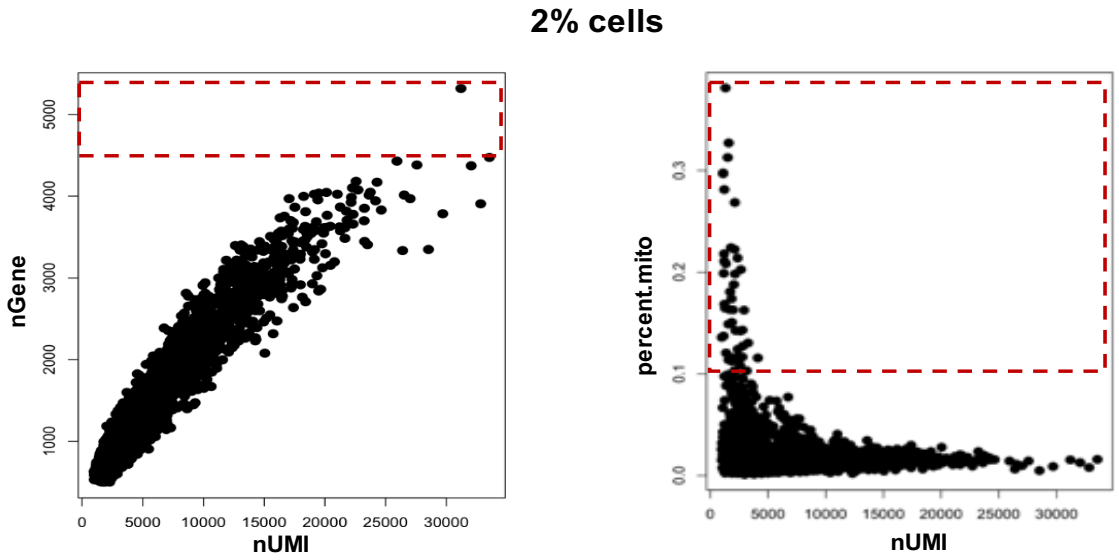
After imposing the filtering criteria described above, I visualized the gene and UMI counts along with the percentage of mitochondrial genes (Fig 4.1) using Seurat v2.4 (Butler *et al.*, 2018) (Methods). The distribution of genes had a range from 500 to 5000 genes in an individual cell with a median count of 1575 genes per cell (Fig 4.1A) for a cut-off of 2% (left), 3% (middle) and 4% (right) of cells. Visualization of UMI counts reflected individual cells with UMI counts having a range from 1000 to 30000 with a median UMI count of 5939 at different cut-offs of 2% (left), 3% (middle) and 4% (right) of cells (Fig 4.1B). After looking at the technical parameters of gene-barcode matrix, I looked at the percentage of mitochondrial genes present in the dataset. An outlier level of mitochondrial content is deterministic of poor biological quality of the cells sequenced. The percentage of mitochondrial content was found to be within 0-0.1% range at different cut-offs of 2% (left), 3% (middle) and 4% (right) of cells (Fig 4.1C), suggesting that very few cells subjected to scRNA-seq were of poor quality.

Having visualized the distribution of gene and UMI counts along with mitochondrial content in the processed matrix, I further studied the relationship of gene counts and UMI counts with respect to mitochondrial content. The plots depicting these relationships aided in defining outlier cells having high mitochondrial gene expression and/or low UMI count. A subset of cells having unique gene counts greater than 4500, which did not follow a linear relationship between UMI and gene count, were excluded for downstream analysis at each cut-off of 2% (Fig 4.2A left), 3% (Fig 4.2B left) and 4% (Fig 4.2C left), as recommended by Seurat authors. These outliers may reflect potential doublets and hence, were filtered out for the downstream analysis. To further exclude any biological artifacts, cells having mitochondrial content of more than 0.1% were filtered out for each cut-off of 2% (Fig 4.2A right), 3% (Fig 4.2B right) and 4% (Fig 4.2C right).

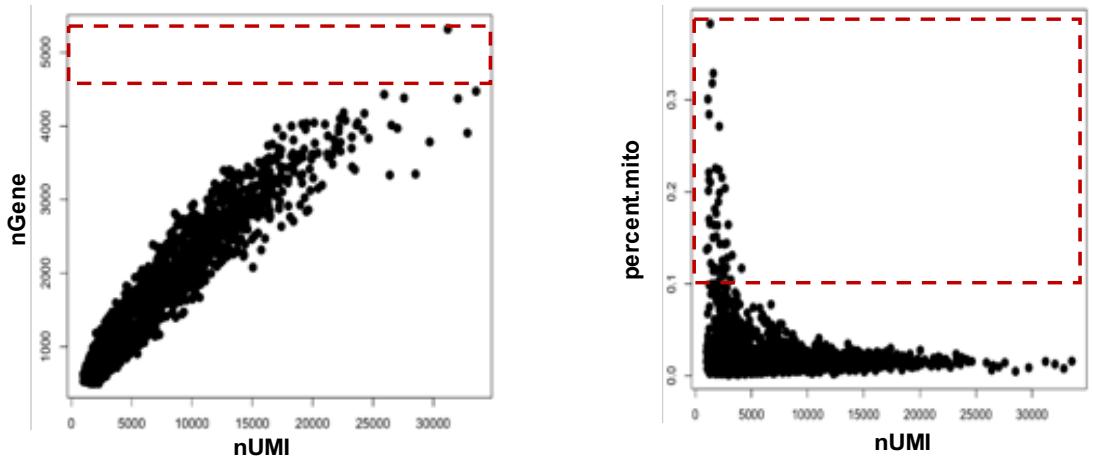
After filtering the outliers, 1739 cells and 1675 cells were obtained at the 2% and 3% cut-offs respectively, whereas the 4% cut-off yielded 1599 cells.

After removing technical and biological artifacts from the dataset, the gene expression measurements were normalized for sequencing depth using the log transformation in Seurat v2.4. This method normalizes individual gene expression values by the total gene expression measurement of a cell and then log-transforms the result after multiplying by a scaling factor of 10,000 by default.

A


2% cells
nGene
nUMI
percent.mito
nUMI

B


3% cells
nGene
nUMI
percent.mito
nUMI

C

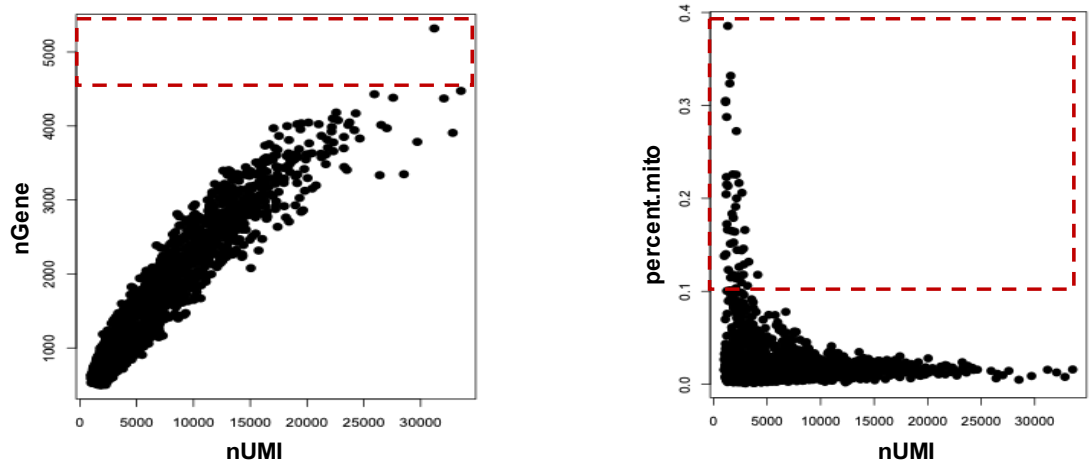
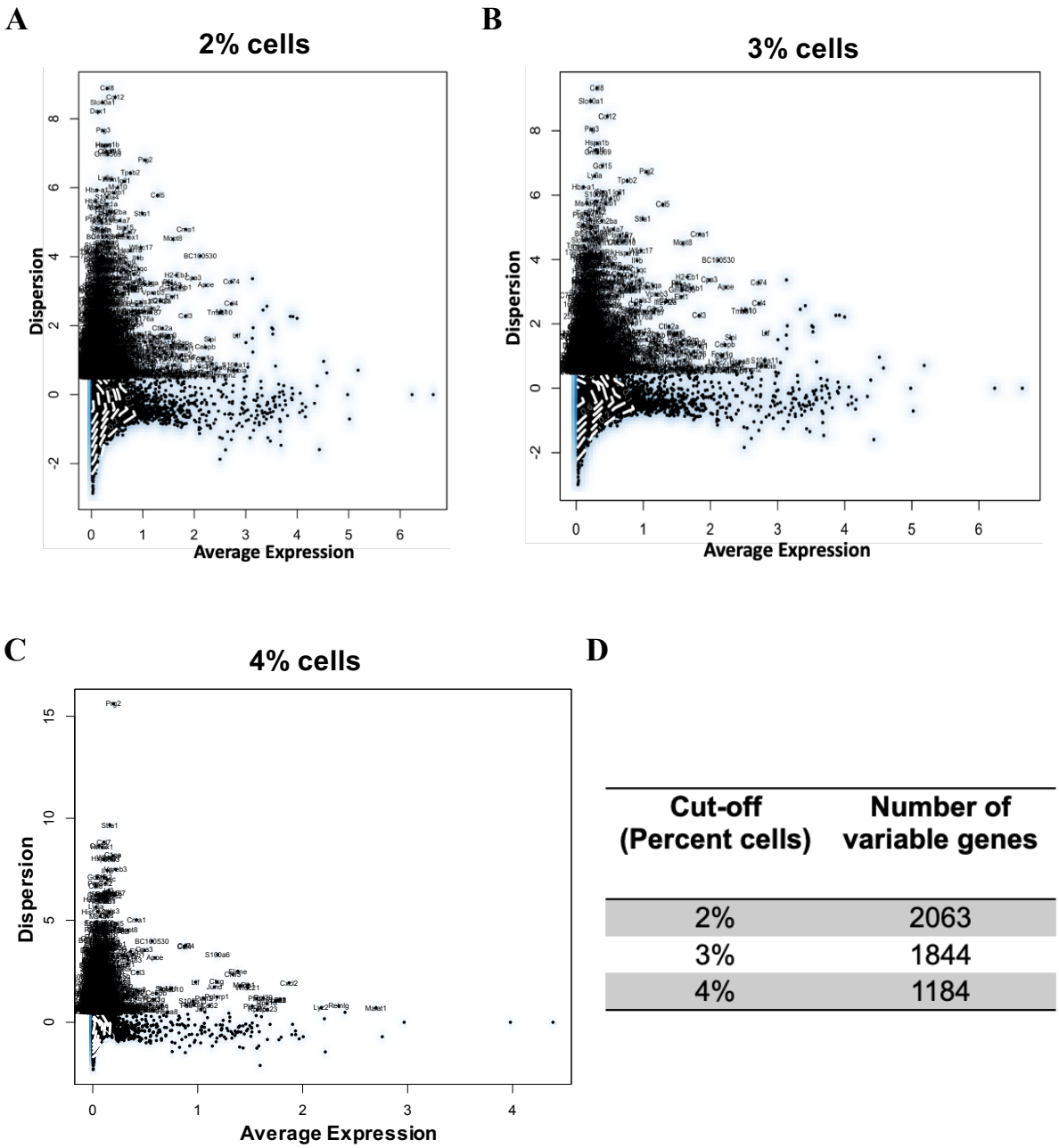

4% cells
nGene
nUMI
percent.mito
nUMI

Figure 4.2: Relationship between UMI counts, gene counts and mitochondrial content using Seurat v2.4.

Visualization of relationships between UMI and gene counts (left), and UMI counts and percentage mitochondrial content to identify and filter outlier cells at different cut-off levels imposed on the percentage of cells that each gene is expressed in **(A)** 2%, **(B)** 3%, **(C)** 4% of cells. A single cut-off of each cell having at least 500 genes was imposed to filter cells. The cells highlighted within the red box were filtered out post analysis.

After filtering and normalization of data to the same depth, highly variable genes were calculated using Seurat v2.4 and analysed further. These highly variable genes strictly define gene–gene similarities or cell–cell similarities, which provide a basis for imputing dropouts with appropriate values. Several pipelines built for scRNA-seq analysis utilize highly variable genes as these are less affected by dropouts, and this approach has been proven effective because the cell-to-cell variability of major phenotypes can often be captured by genes exhibiting high variability (Qiu, 2020). For this, each gene was placed in a bin and a z-score was calculated for individual bin to plot a relationship between average expression and dispersion (Fig 4.3). These plots were created for the three different cut-offs of 2% (Fig 4.3A), 3% (Fig 4.3B) and 4% (Fig 4.3C) and 2063, 1844, and 1184 highly variable genes were identified for the respective cut-offs (Fig 4.3D). Since Seurat aims to identify nearly 2000 highly variable genes at a given cut-off, the output of 1184 variable genes at 4% cut-off indicated that was a stringent cut-off which may lead to the filtering out of rare cell populations.



Cut-off (Percent cells)	Number of variable genes
2%	2063
3%	1844
4%	1184

Figure 4.3: Detection of variable genes across single cells.

Plot between dispersion and average expression of individual genes based on z-score of dispersion for each bin, where genes are divided into bins based on their mean expression. This was done for different cut-offs for the percentage of cells in which a gene expressed **(A)** 2%, **(B)** 3%, **(C)** 4% of cells. Cells were filtered using a single cut-off of having at least 500 genes, **(D)** Tabular representation of number of variable genes obtained at individual cut-offs imposed to detect variable genes in the given dataset.

4.3 Dimensionality reduction analysis

4.3.1 Principal Component Analysis (PCA)

After filtering, normalization, and detection of sources of variability in the sequenced data, the next step was to visualize the data in a reduced dimensional space. Prior to that, it was important to remove unwanted sources of technical and biological noise to improve downstream analysis. For doing this, I made use of the ScaleData function implemented in Seurat v2.4, which constructs linear models of gene expression estimation and generates a z-scale score for the same. This regresses out noisy signals from the data. Then, I performed linear dimensionality reduction analysis using Principal Component Analysis (PCA) taking variable genes found in Figure 4.3 as an input for individual cut-off. Representative PCA plots based on scores for PC1 and PC2 are shown in Figure 4.4 where (A), (C) and (E) represent PC plots for cut-off of 2%, 3% and 4% respectively. To overcome the extensive technical noise associated with a particular gene in the dataset, the cells were visualized using a clustering method which employed PC scores as a correlation measure. To determine how many PCs to be included in downstream analysis, Elbow plot analysis (Cattell RB, 1966) was done. The analysis took standard deviation of each PC into consideration to determine the significant number of PCs. In case of 2% cut- off, there were 12 significant PCs (Fig 4.4B) whereas in case of 3% and 4% cut-off, the number of significant PCs were 10 and 8 respectively (Fig 4.4D and 4.4F). Another approach known as the Jackstraw plot approach was followed to determine the significant number of PCs. The JackStrawPlot function provided a visualization tool for comparing the distribution of p-values for each PC against a uniform distribution (dashed line). Significant PCs showed a strong enrichment of genes with low p-values (solid curve above the dashed line). Based on jackstraw approach I found all 12 PCs to be significant for all individual cut-offs including 2% (Fig 4.5A), 3% (Fig 4.5B), and 4% (Fig 4.5C). In order to study the association of sets of genes with respective PCs, VizPCA function implemented in Seurat was utilised, which allowed easy exploration of genes which tend to be the primary source of variability within the dataset for cut-off of 2% (Fig 4.6A), 3% (Fig 4.6B) and 4% (Fig 4.6C).

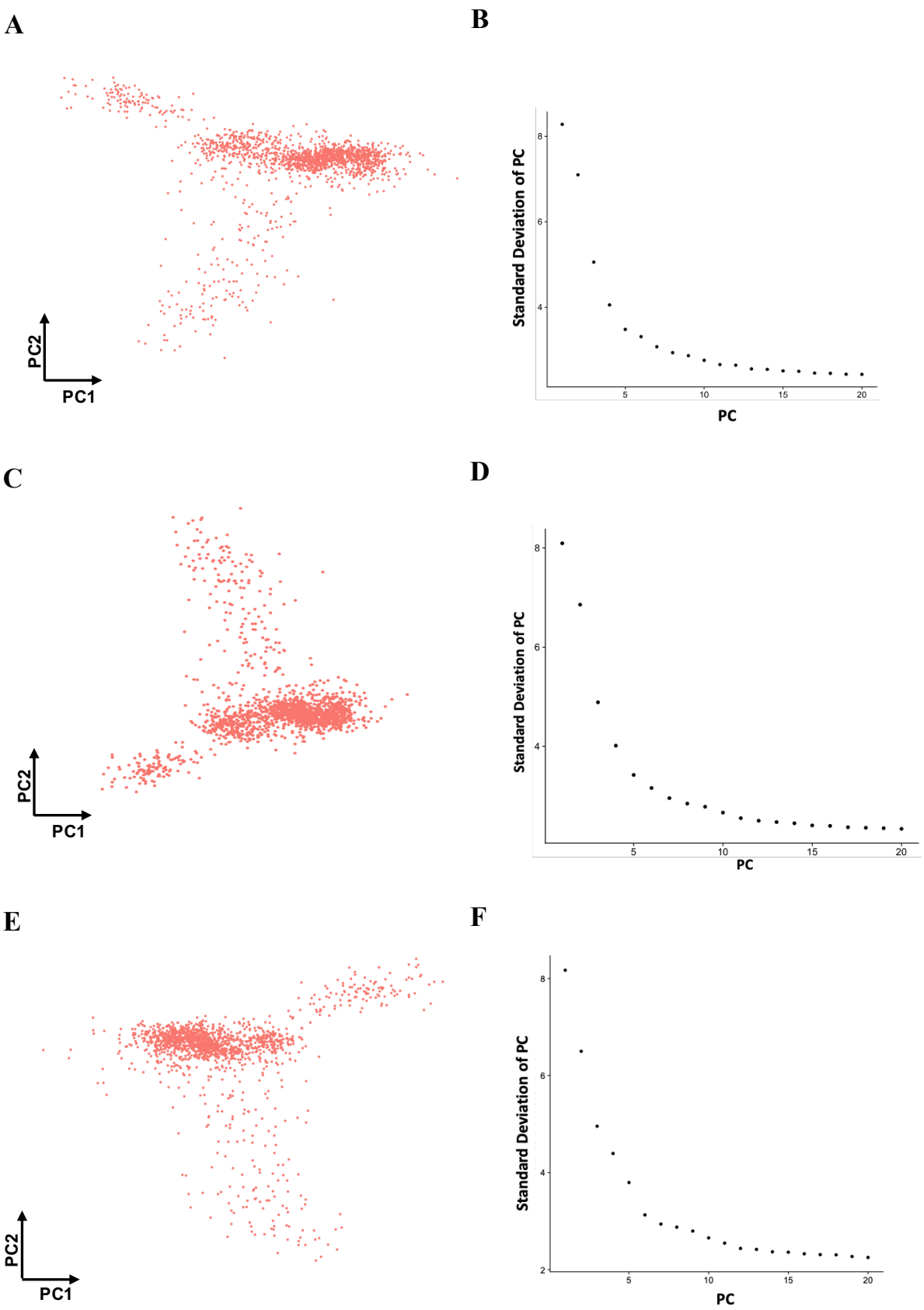
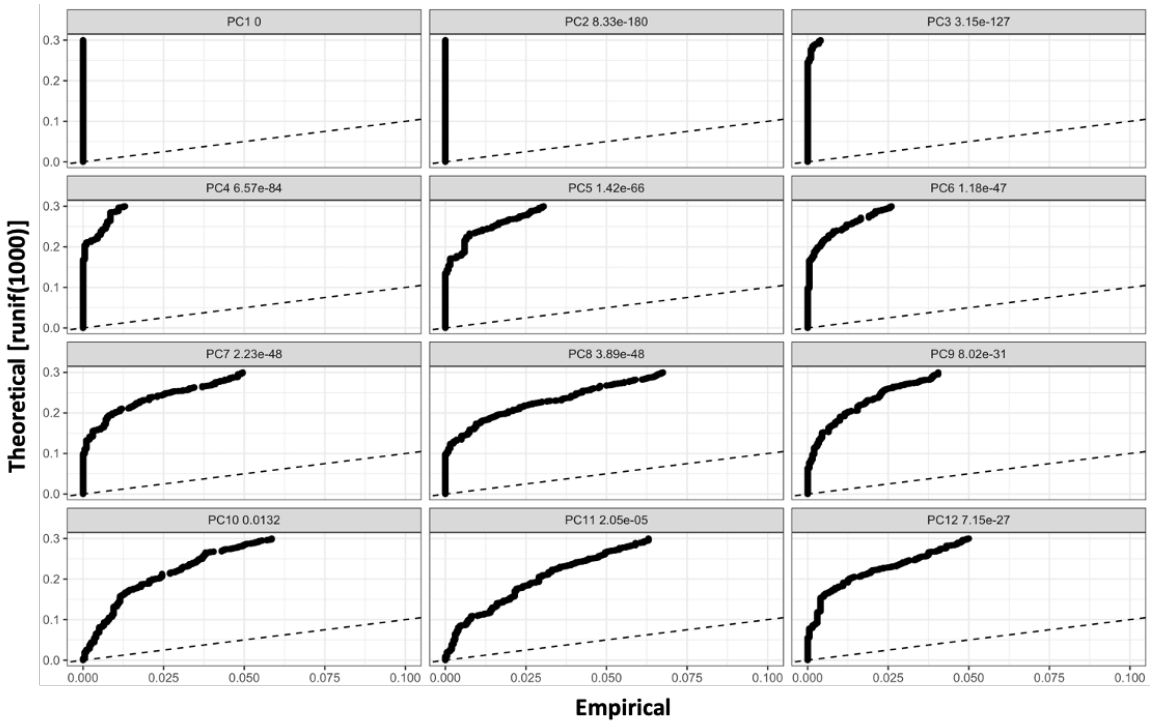
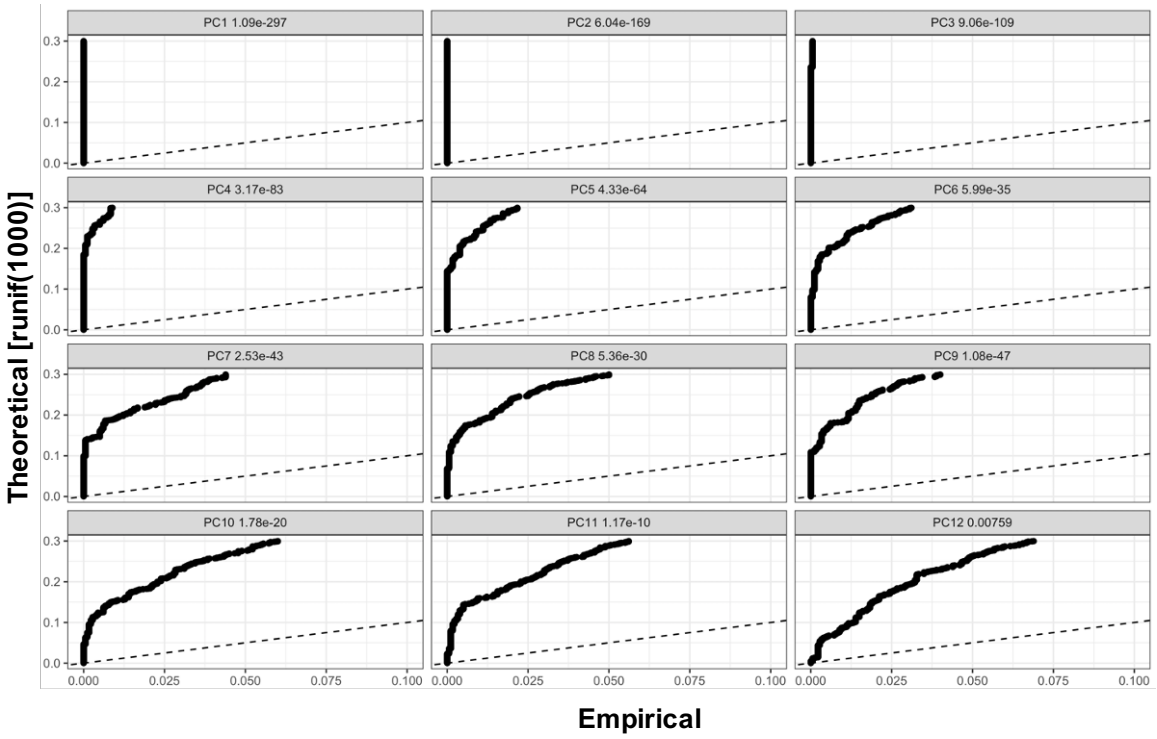

A
PC2
PC1
B
Standard Deviation of PC
PC
C
PC2
PC1
D
Standard Deviation of PC
PC
E
PC2
PC1
F
Standard Deviation of PC
PC

Figure 4.4: Dimensionality reduction using Principal Component Analysis.
Principal Component analysis in order to visualize data in lower dimensional space (left) and PC Elbow plot for identification of significant PCs based on their standard deviation (right) for different cut-offs of **(A) and (B)** 2%, **(C) and (D)** 3%, **(E) and (F)** 4% of cells expressing a particular gene.

A


PC1 0
PC2 8.33e-180
PC3 3.15e-127
PC4 6.57e-84
PC5 1.42e-66
PC6 1.18e-47
PC7 2.23e-48
PC8 3.89e-48
PC9 8.02e-31
PC10 0.0132
PC11 2.05e-05
PC12 7.15e-27
Theoretical [runif(1000)]
Empirical

B


PC1 1.09e-297
PC2 6.04e-169
PC3 9.06e-109
PC4 3.17e-83
PC5 4.33e-64
PC6 5.99e-35
PC7 2.53e-43
PC8 5.36e-30
PC9 1.08e-47
PC10 1.78e-20
PC11 1.17e-10
PC12 0.00759
Theoretical [runif(1000)]
Empirical

C

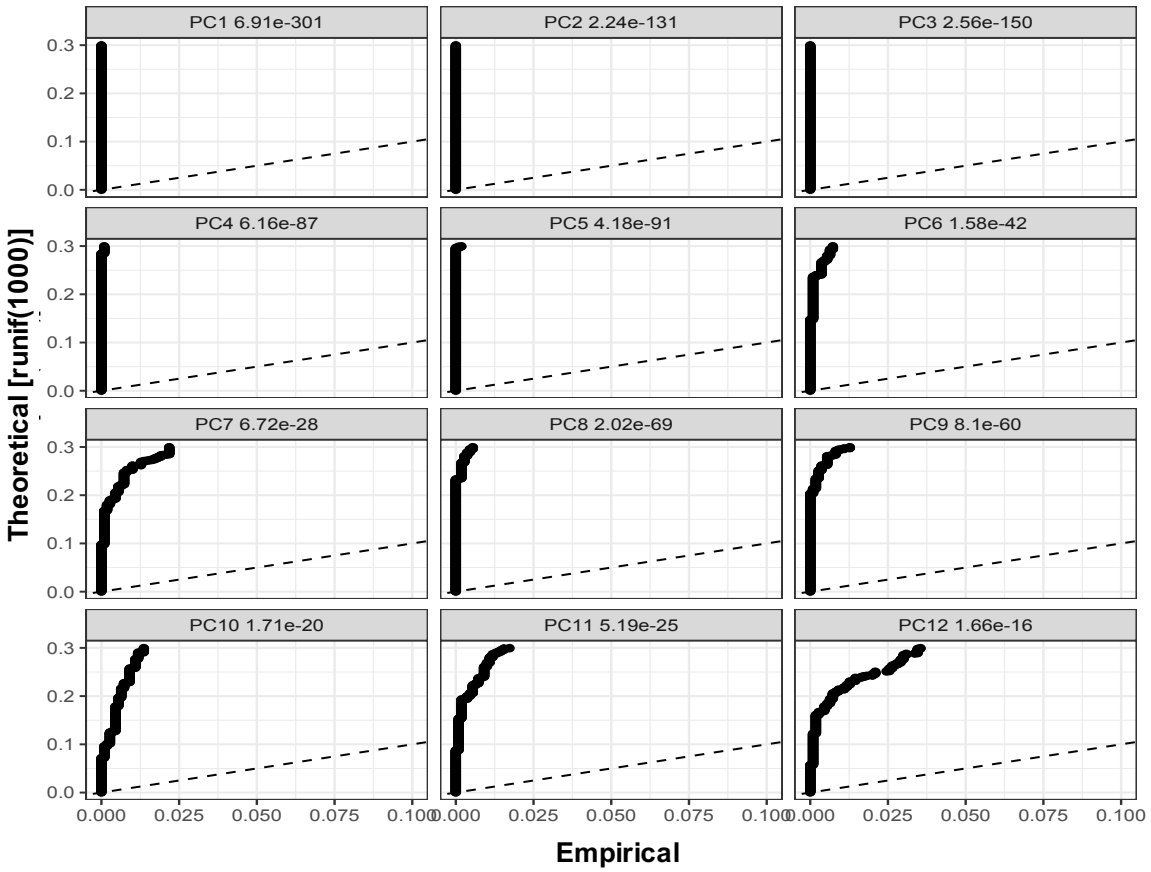
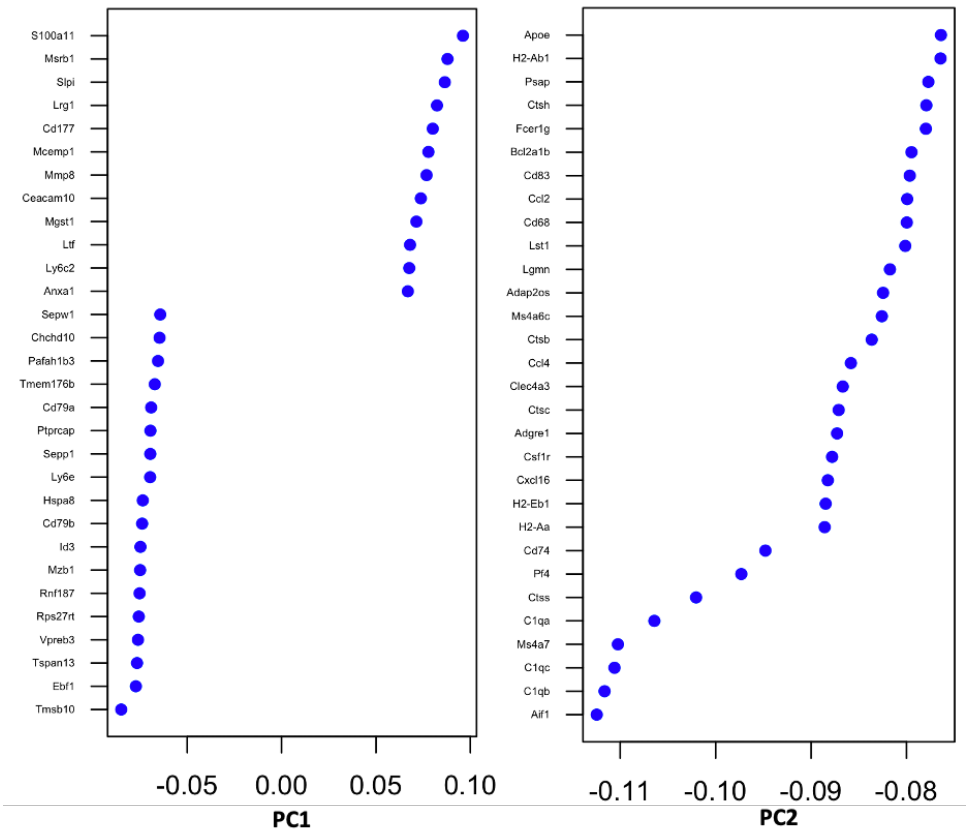


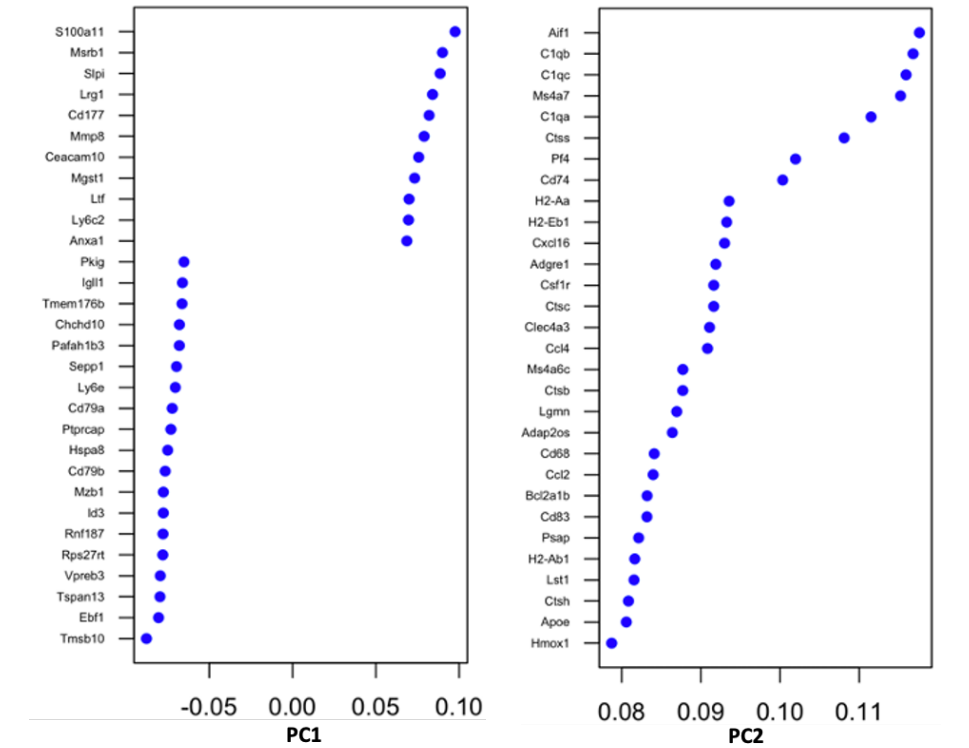
Figure 4.5: Jackstraw plot for identification of significant PCs.

Jackstraw plot for comparing the distribution of p-values for each PC against a uniform distribution (dashed line). Significant PCs showed a strong enrichment of genes with low p-values (solid curve above the dashed line), **(A)** Jackstraw plot for 2% cut-off, **(B)** Jackstraw plot for 3% cut-off and **(C)** Jackstraw plot for 4% cut-off.

A


S100a11
Msrb1
Slpi
Lrg1
Cd177
Mcemp1
Mmp8
Ceacam10
Mgst1
Ltf
Ly6c2
Anxa1
Sepw1
Chchd10
Pafah1b3
Tmem176b
Cd79a
Ptprcap
Sepp1
Ly6e
Hspa8
Cd79b
Id3
Mzb1
Rnf187
Rps27rt
Vpreb3
Tspan13
Ebf1
Tmsb10
-0.05
0.00
0.05
0.10
PC1
Apoe
H2-Ab1
Psap
Ctsh
Fcer1g
Bcl2a1b
Cd83
Ccl2
Cd68
Lst1
Lgmn
Adap2os
Ms4a6c
Ctsb
Ccl4
Clec4a3
Ctsc
Adgre1
Csf1r
Cxcl16
H2-Eb1
H2-Aa
Cd74
Pf4
Ctss
C1qa
Ms4a7
C1qc
C1qb
Aif1
-0.11
-0.10
-0.09
-0.08
PC2

B


S100a11
Msrb1
Slpi
Lrg1
Cd177
Mmp8
Ceacam10
Mgst1
Ltf
Ly6c2
Anxa1
Pkig
Igll1
Tmem176b
Chchd10
Pafah1b3
Sepp1
Ly6e
Cd79a
Ptprcap
Hspa8
Cd79b
Mzb1
Id3
Rnf187
Rps27rt
Vpreb3
Tspan13
Ebf1
Tmsb10
-0.05
0.00
0.05
0.10
PC1
Aif1
C1qb
C1qc
Ms4a7
C1qa
Ctss
Pf4
Cd74
H2-Aa
H2-Eb1
Cxcl16
Adgre1
Csf1r
Ctsc
Clec4a3
Ccl4
Ms4a6c
Ctsb
Lgmn
Adap2os
Cd68
Ccl2
Bcl2a1b
Cd83
Psap
H2-Ab1
Lst1
Ctsh
Apoe
Hmox1
0.08
0.09
0.10
0.11
PC2

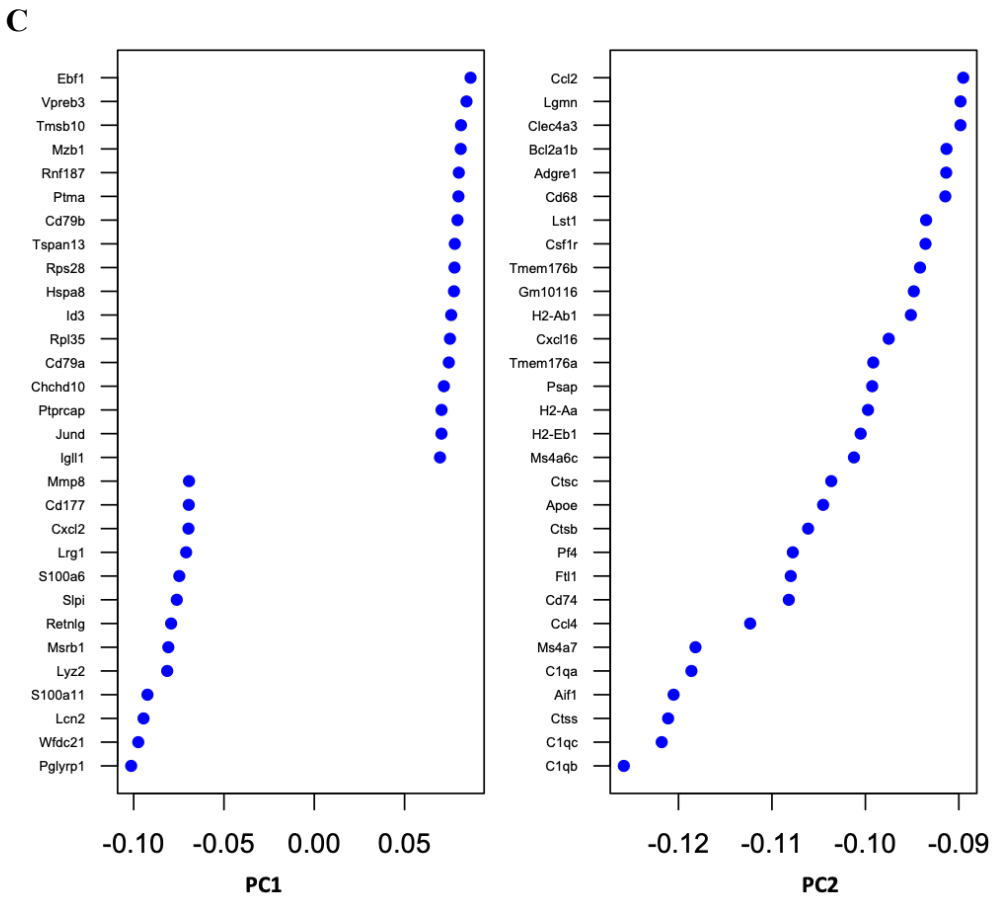


Figure 4.6: Visualization of PCs.

Genes associated with PC1 and PC2 for (A) 2% cut-off, (B) 3% cut-off, and (C) 4% cut-off.

4.3.2 t-Distributed Stochastic Neighbour Embedding (t-SNE)

After determining significant PCs using Elbow plot, I looked at how these PCs may aid in visualization of data by separating the cells into different clusters. For this, graph- based clustering approach implemented in Seurat v2.4 was followed, based on above defined PCs. Briefly, the graph-based clustering approach followed K-nearest neighbour algorithm (KNN) (Sieranoja, 2018), where edges were drawn between cells with similar gene expression patterns and grouped based on a modularity optimization technique, the Louvain algorithm (Blondel, 2008), which iteratively grouped cells together. In order to lay over the clusters obtained in a low dimensional space, t-Distributed Stochastic Neighbour Embedding (tSNE) tool (Van Der Maaten and Hinton, 2008) was used which places cells with similar local neighbourhoods in

the high-dimensional space together in the low-dimensional space (Fig 4.7). I obtained 5 clusters, with less defined boundaries in low dimensional space for the 2% cut-off (Fig 4.7A), whereas well-segregated clusters were obtained for 3% (6 clusters, Fig 4.7B) and 4% cut-offs (8 clusters, Fig 4.7C). Although both 3% and 4% cut-off led to well defined clusters of cells in lower dimensional space, however, 4% cut-off had much lower number of cells (1599 cells in 4% compared to 1675 cells in 3%) indicative of really strict filtering. Therefore, I utilized the gene-barcode matrix obtained upon applying the 3% cut-off for further downstream analysis.

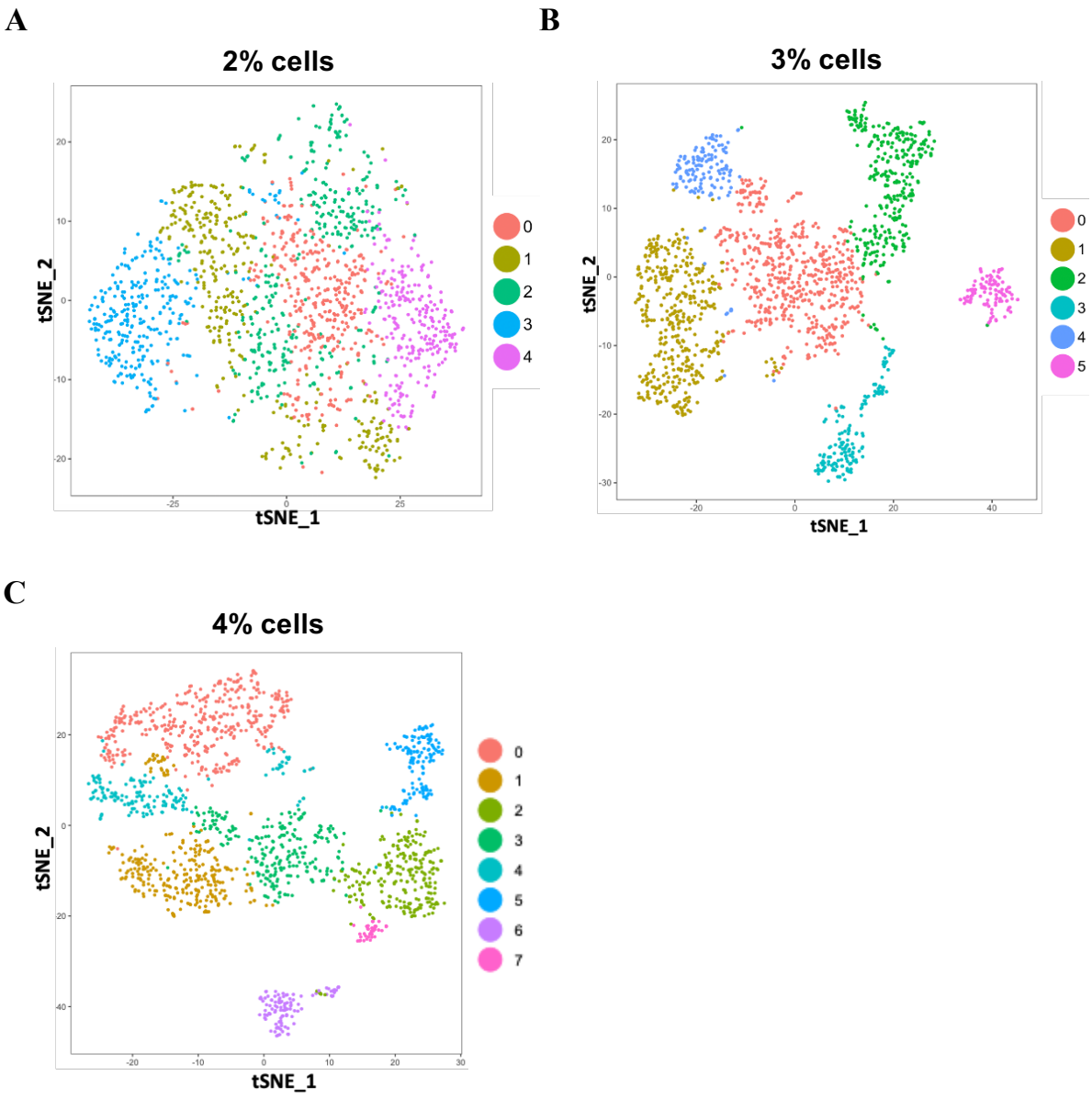


Figure 4.7: Dimensionality reduction analysis using t-distributed stochastic neighbouring method.

tSNE representation of graph-based clusters obtained based on significant PCs for different cut-offs where each gene is expressed in at least, **(A)** 2%, **(B)** 3%, **(C)** 4% of cells that express a minimum of 500 genes.

4.4 *Kat2a* loss leads to global downregulation of gene expression

After initial pre-processing steps including filtering and normalization, the gene-barcode matrix obtained had 1675 cells; 361 cells from *Kat2a* WT at 2 months, 331 cells from *Kat2a* NULL at 2 months, 502 cells from *Kat2a* WT at 4 months, and 481 cells from *Kat2a* NULL at 4 months, representing transcript levels of 8603 genes normalized to the same sequencing depth. These cells had a median gene count of 1575 and median UMI count of 5939 (Table 4.1).

Table 4.1: Data summary post pre-processing and filtering steps

Number of cells sequenced	1739
Number of cells post-filtering	1675
Kat2a WT-2 Months	361
Kat2a NULL-2 Months	331
Kat2a WT-4 Months	502
Kat2a NULL-4 Months	481
Median genes per cell	1575
Median UMI counts per cell	5939

To study the effect of *Kat2a* loss on global gene expression, I performed differential expression analysis to identify genes expressed differentially upon loss of *Kat2a* at different time points. For this, DESeq2 method (Love, Huber and Anders, 2014) integrated in Seurat v2.4 was employed which identifies differentially expressed genes based on a negative binomial distribution (Methods). I started by comparing *Kat2a* WT cells from both 2 months and 4 months time-points against *Kat2a* NULL cells from both time points. This global comparison revealed that *Kat2a* NULL leads to an overall downregulation of gene expression, with 811 genes (p-adj <0.05) being downregulated compared to 70 genes being upregulated (p-adj <0.05) (Table 4.2). Interestingly, the global comparison also suggested that only 82 (7+75) genes out of 881 (811+70) displayed a higher than 20% change in gene expression upon *Kat2a* loss, suggesting that loss of *Kat2a* may be contributing towards transcriptional instability of gene expression programmes during pre-leukaemia progression rather than a mere downregulation of genes as evident from the $\log_2$ fold change differences.

Table 4.2: DESeq2 analysis for different comparisons with numbers representing upregulated and downregulated genes with p-adj <0.05. Numbers inside the brackets represent minimum 20% fold change difference

Comparison	Upregulated genes	Downregulated genes
Kat2a NULL vs *Kat2a* WT Total	70 (7)	811 (75)
Kat2a NULL vs *Kat2a* WT 2M	48 (44)	551 (197)
Kat2a NULL vs *Kat2a* WT 4M	39 (38)	118 (73)
Kat2a WT 4M vs 2M	21 (21)	600 (115)
Kat2a NULL 4M vs 2M	79 (37)	1325 (353)
Kat2a NULL 2M vs *Kat2a* WT 4M	123 (94)	493 (172)

After performing the global comparison, I compared gene expression in *Kat2a* NULL cells against *Kat2a* WT cells at individual time points, where the comparison between 2 months

samples would indicate *Kat2a* associated transcriptional changes during pre-leukaemia initiation and the comparison between 4 months samples would indicate the transcriptional changes associated with *Kat2a* loss during pre-leukaemia maintenance. Downregulated genes upon loss of *Kat2a* at each time point again outnumbered the upregulated genes, consistent with the global comparison (Table 4.2). There were 551 downregulated and 48 upregulated genes upon loss of *Kat2a* at 2 months, whereas 118 genes were downregulated and 39 genes were upregulated upon loss of *Kat2a* at 4 months post transplantation. This indicated that *Kat2a* loss may potentially impact gene expression during early stages of pre-leukaemia transformation (Table 4.2). Comparing the genes downregulated in the global as well as the respective time point comparisons, only 52 genes displayed consistent downregulation (Fig 4.8A). Next, I compared the transcriptional changes in *Kat2a* WT occurring between 2 months and 4 months and *Kat2a* NULL occurring between 2 months and 4 months separately. *Kat2a* WT cells showed a downregulation of 600 genes at 4 months compared to 2 months, whereas 1325 genes were downregulated in *Kat2a* NULL cells at 4 months compared to 2 months (Table 4.2). The higher number of downregulated genes in *Kat2a* NULL time-series comparison may indicate that *Kat2a* loss leads to downregulation of a specific set of genes during the process of pre-leukaemia progression and further suggests that downregulation of gene expression is a hallmark of pre-leukaemia progression. Only 38 genes were common to the set of genes downregulated in time series comparison and the set of genes downregulated in the global comparison (Fig 4.8B).

Since the combined *in vitro* and *in vivo* analysis suggested an accelerated *RUNX1-RUNX1T1(9a)* leukaemia progression upon *Kat2a* loss (Chapter-3), I wanted to compare the leukaemia progression hierarchy in *Kat2a* WT cells at 4 months and *Kat2a* NULL cells at 2 months. For this, I looked at differentially expressed genes in *Kat2a* NULL cells at 2 months post transplantation compared to *Kat2a* WT cells at 4 months post transplantation. There were 493 genes downregulated and 123 genes upregulated in *Kat2a* NULL at 2 months post transplantation highlighting gene expression programmes specific to *Kat2a* loss during the pre-leukaemia transformation process (Table 4.2).

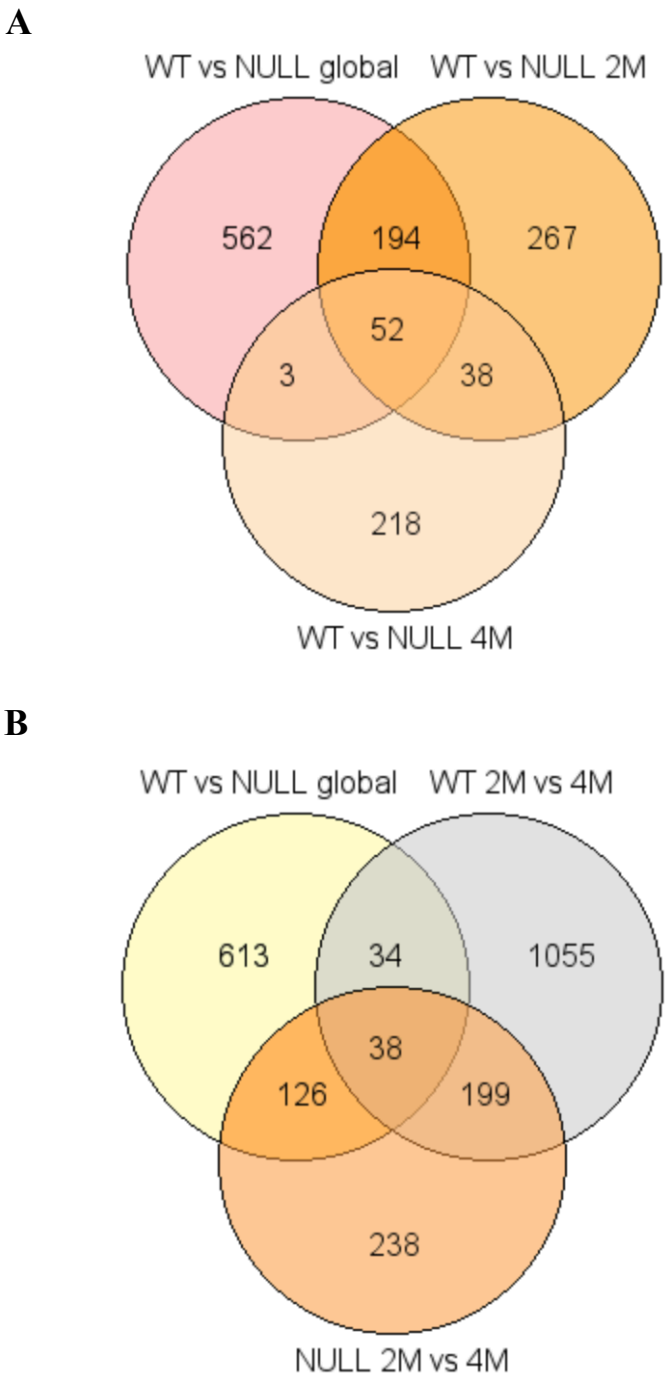


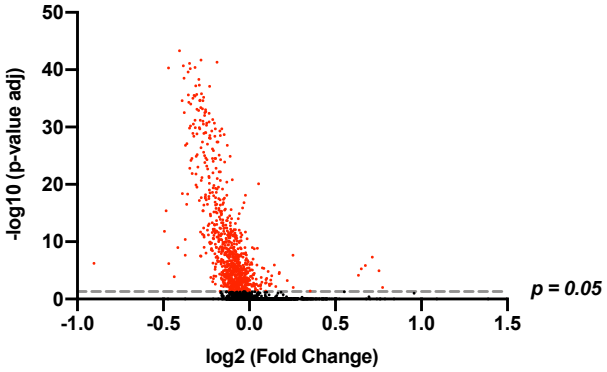
Figure 4.8: Venn diagram for common set of downregulated genes.

(A) Venn diagram highlighting common set of genes among global *Kat2a* WT and *Kat2a* NULL differential comparison along with the *Kat2a* WT vs NULL comparison at respective time points, **(B)** Venn diagram highlighting common set of genes among global *Kat2a* WT and *Kat2a* NULL differential comparison along with time point comparisons for respective genotypes.

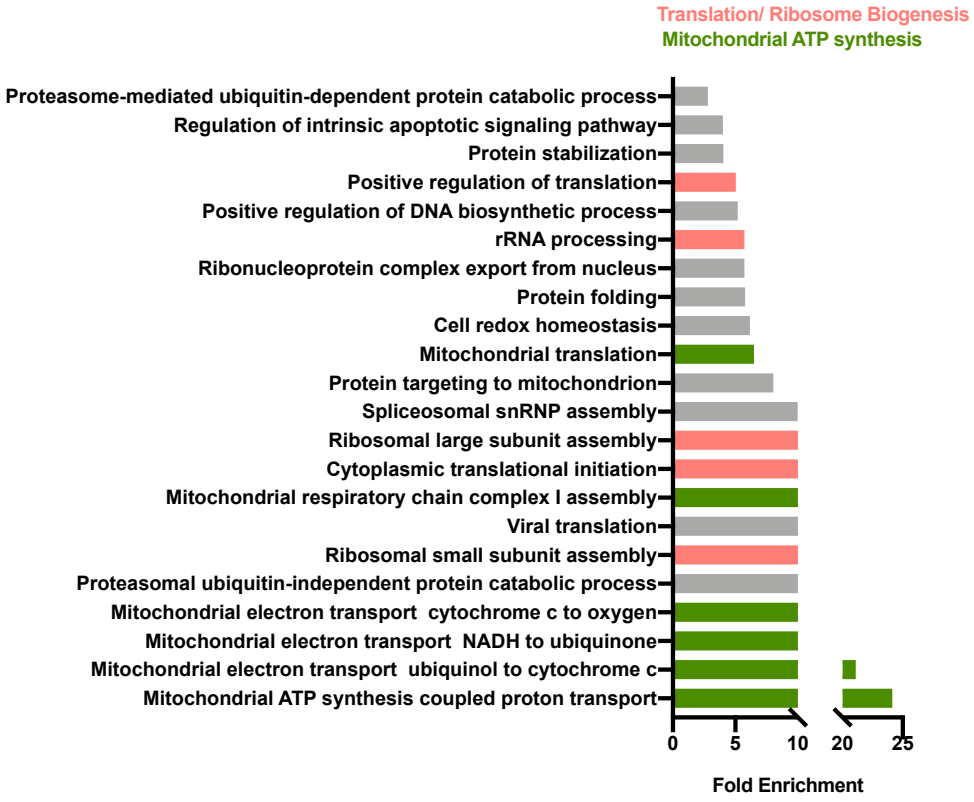
4.5 The downregulated genes were enriched in mitochondrial ATP synthesis, ribosomal biogenesis and cytoplasmic translation pathways

Having observed that loss of *Kat2a* leads to global downregulation of gene expression, I next wanted to study the gene expression pathways enriched upon *Kat2a* loss. For this, I started with the list of genes which were significantly downregulated upon loss of *Kat2a* in a global comparison, using the DESeq2 method mentioned previously (Fig 4.9A). The 811 genes which were downregulated in *Kat2a* NULL cells, were subjected to Panther (version 14.0) enrichment analysis (H. Mi *et al.*, 2018) (Methods). The two major enriched categories were associated with translation/ribosome biogenesis and mitochondrial ATP synthesis pathways (Fig 4.9B). Some of the other enriched categories included protein stabilization, ubiquitination, splicing, protein folding, apoptosis etc., highlighting that loss of *Kat2a* may exert its effect through general pathways rather than leukaemia specific ones. This observation was in line with a previous study conducted in our lab (Domingues *et al.*, 2020). I used another approach to conduct pathway enrichment analysis, specifically, by overlapping these downregulated genes with MSigDB (Subramanian *et al.*, 2005; Liberzon *et al.*, 2015) (Methods). The top 10 gene sets with FDR $q < 0.05$ were enriched in translational activity/ribosomal structure and mitochondrial activity, consistent with the results obtained from Panther enrichment analysis (Fig 4.9C).

A


50
40
30
20
10
0
-log10 (p-value adj)
p = 0.05
-1.0
-0.5
0.0
0.5
1.0
1.5
log2 (Fold Change)

B


Translation/ Ribosome Biogenesis
Mitochondrial ATP synthesis
Proteasome-mediated ubiquitin-dependent protein catabolic process
Regulation of intrinsic apoptotic signaling pathway
Protein stabilization
Positive regulation of translation
Positive regulation of DNA biosynthetic process
rRNA processing
Ribonucleoprotein complex export from nucleus
Protein folding
Cell redox homeostasis
Mitochondrial translation
Protein targeting to mitochondrion
Spliceosomal snRNP assembly
Ribosomal large subunit assembly
Cytoplasmic translational initiation
Mitochondrial respiratory chain complex I assembly
Viral translation
Ribosomal small subunit assembly
Proteasomal ubiquitin-independent protein catabolic process
Mitochondrial electron transport cytochrome c to oxygen
Mitochondrial electron transport NADH to ubiquinone
Mitochondrial electron transport ubiquinol to cytochrome c
Mitochondrial ATP synthesis coupled proton transport
0
5
10
20
25
Fold Enrichment

C

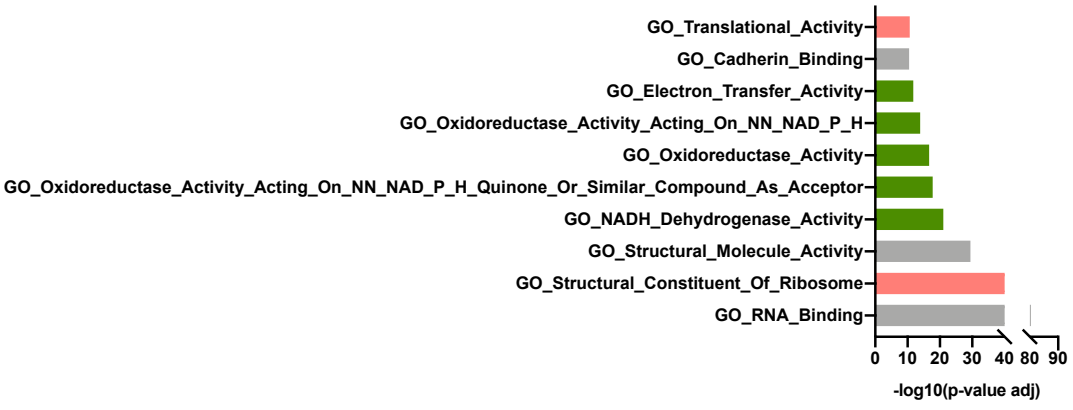
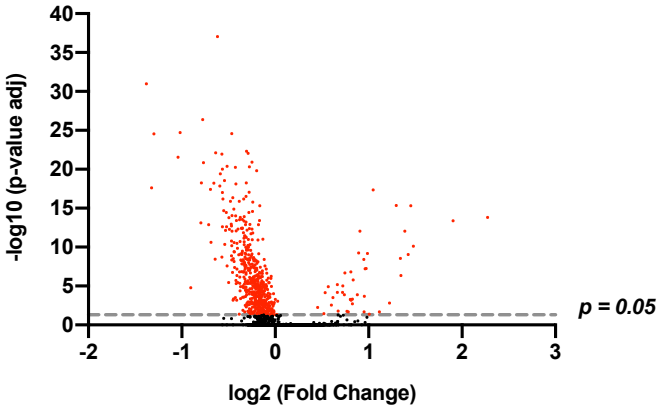

GO_Translational_Activity
GO_Cadherin_Binding
GO_Electron_Transfer_Activity
GO_Oxidoreductase_Activity_Acting_On_NN_NAD_P_H
GO_Oxidoreductase_Activity
GO_Oxidoreductase_Activity_Acting_On_NN_NAD_P_H_Quinone_Or_Similar_Compound_As_Acceptor
GO_NADH_Dehydrogenase_Activity
GO_Structural_Molecule_Activity
GO_Structural_Constituent_Of_Ribosome
GO_RNA_Binding
0
10
20
30
40
80
90
-log10(p-value adj)

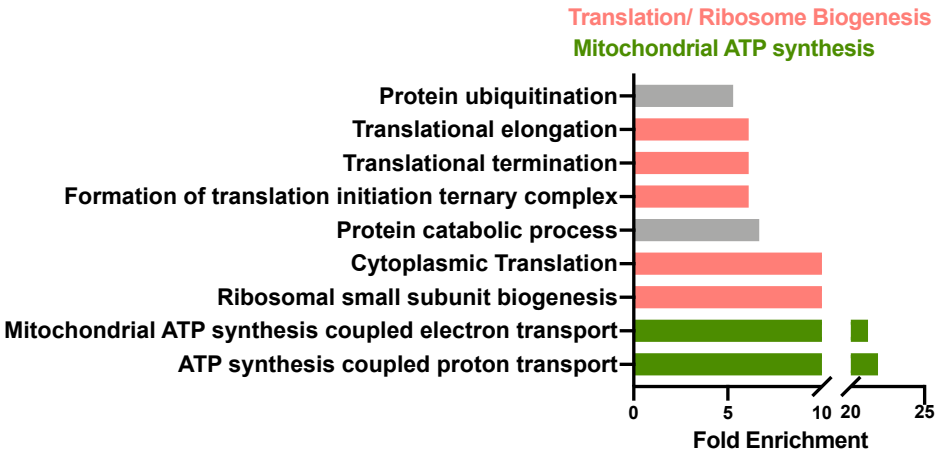
Figure 4.9: DESeq2 analysis for *Kat2a* WT vs *Kat2a* NULL global comparison.
(A) Volcano plot representing differentially expressed genes obtained from comparing *Kat2a* WT (2 months + 4 months) vs *Kat2a* NULL (2 months + 4 months) cells. The genes present above the grey line (p-adj=0.05) are significant and highlighted in red, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-adj <0.05) based on Fisher's exact test with Bonferroni correction, where pink highlights genes associated with translation/ribosome biogenesis, and green highlights mitochondrial ATP synthesis associated genes, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

After performing pathway enrichment analysis for the global *Kat2a* NULL versus *Kat2a* WT comparison, I performed similar analysis for genes downregulated at individual time-points. The comparison between 2 months samples would indicate *Kat2a* associated gene expression pathways affected during pre-leukaemia initiation whereas comparison between 4 months samples would indicate the transcriptional programmes associated with *Kat2a* loss during pre-leukaemia maintenance. There were 599 significantly differentially expressed genes in *Kat2a* NULL compared to *Kat2a* WT at 2 months post transplantation, out of which 551 genes were downregulated (Fig 4.10A). Gene ontology analysis using Panther indicated enrichment of similar gene expression programmes as global comparisons, namely translation/ribosomal biogenesis and mitochondrial ATP synthesis (Fig 4.10B). However, unlike the global comparison, this analysis did not capture other protein modification processes, highlighting the importance of these 2 pathways during leukaemia initiation upon *Kat2a* loss. This observation was consistent with MSigDB overlap where translation/ribosomal biogenesis and mitochondrial ATP synthesis pathways were prominently enriched (Fig 4.10C).

A


-log10 (p-value adj)
40
35
30
25
20
15
10
5
0
-2
-1
0
1
2
3
log2 (Fold Change)
p = 0.05

B


Translation/ Ribosome Biogenesis
Mitochondrial ATP synthesis
Protein ubiquitination
Translational elongation
Translational termination
Formation of translation initiation ternary complex
Protein catabolic process
Cytoplasmic Translation
Ribosomal small subunit biogenesis
Mitochondrial ATP synthesis coupled electron transport
ATP synthesis coupled proton transport
0
5
10
20
25
Fold Enrichment

C

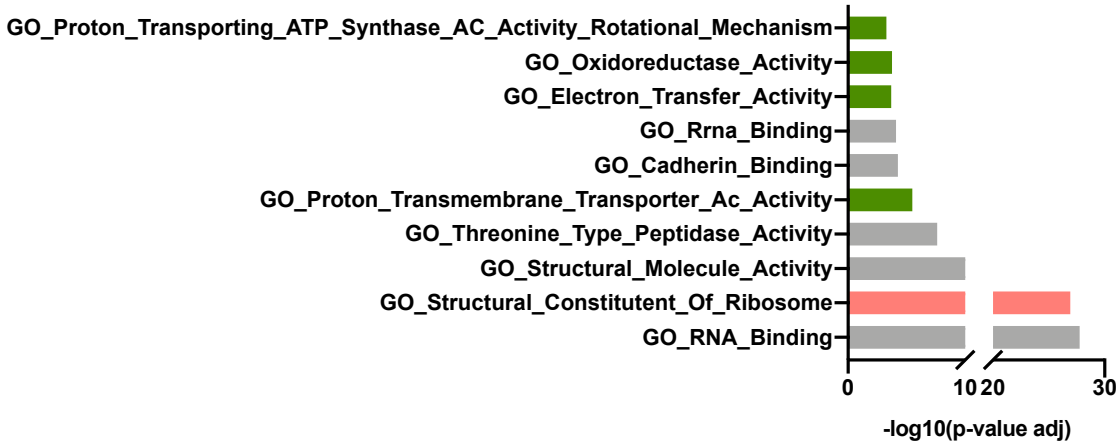
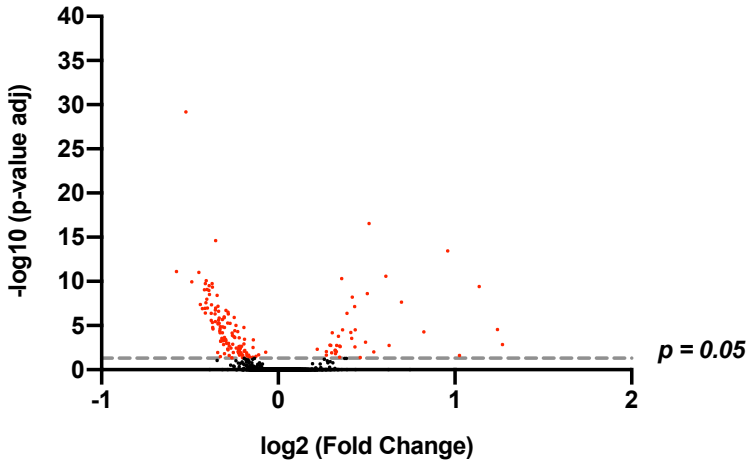

GO_Proton_Transporting_ATP_Synthase_AC_Activity_Rotational_Mechanism
GO_Oxidoreductase_Activity
GO_Electron_Transfer_Activity
GO_Rrna_Binding
GO_Cadherin_Binding
GO_Proton_Transmembrane_Transporter_Ac_Activity
GO_Threonine_Type_Peptidase_Activity
GO_Structural_Molecule_Activity
GO_Structural_Constitutent_Of_Ribosome
GO_RNA_Binding
0
10
20
30
-log10(p-value adj)

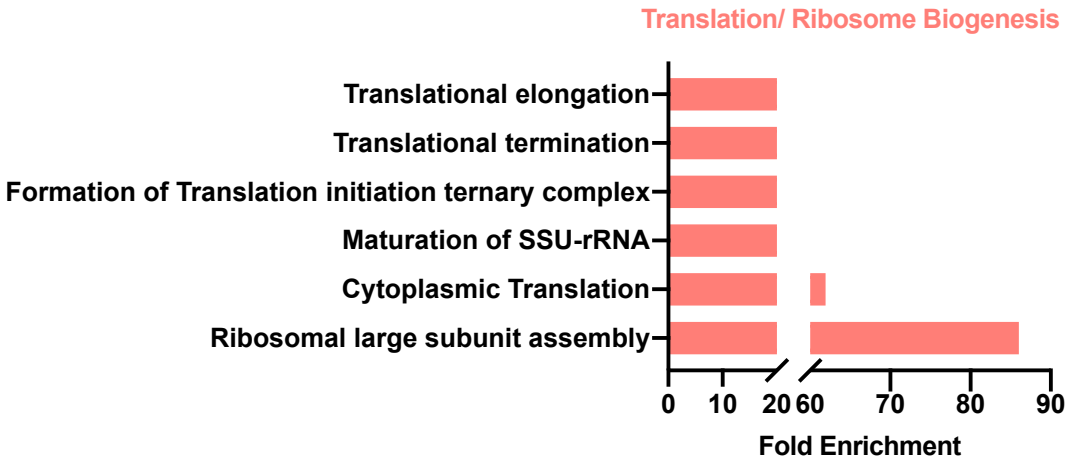
Figure 4.10: DESeq2 analysis for *Kat2a* WT vs *Kat2a* NULL 2months comparison.
(A) Volcano plot representing differentially expressed genes obtained from comparing *Kat2a* WT vs *Kat2a* NULL cells (2 months). The genes present above the grey line (p=0.05) are significant and highlighted in red, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-adj<0.05) based on Fisher's exact test with Bonferroni correction, where pink highlights genes associated with translation/ribosome biogenesis and green highlights mitochondrial ATP synthesis associated genes, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

Further, I looked at the pathways enriched in *Kat2a* NULL at 4 months post transplantation. There were 157 genes in total which were differentially expressed at 4 months (Fig 4.11A), a number much lower than the 599 differentially expressed genes at 2 months, indicating that loss of *Kat2a* leads to transcriptional changes at early stages of *RUNX1-RUNX1T1(9a)* leukaemia transformation. The 118 downregulated genes out of the 157 genes were subjected to Panther gene ontology enrichment analysis, which suggested overrepresentation of genes related to translation and ribosome biogenesis (Fig 4.11B). Unlike previous comparisons performed globally and at 2 months, no other categories were enriched, perhaps indicating that those pathways were specifically required at early stages of transformation whereas translation/ribosome biogenesis were continually required for pre-leukaemia maintenance. This was consistent with previous lab observations (Domingues *et al.*, 2020). However, this may also indicate that the time point at 4 months post transplantation may be too late to capture alterations in the pathways which were enriched globally and at 2 months. Further, I looked at the enrichment of gene expression pathways using the MSigDB overlap approach. The findings were in-line with the Panther gene ontology analysis with enrichment in ribosomal structural biogenesis genes (Fig 4.11C).

A


40
35
30
25
20
15
10
5
0
-log10 (p-value adj)
p = 0.05
-1
0
1
2
log2 (Fold Change)

B


Translation/ Ribosome Biogenesis
Translational elongation
Translational termination
Formation of Translation initiation ternary complex
Maturation of SSU-rRNA
Cytoplasmic Translation
Ribosomal large subunit assembly
0
10
20
60
70
80
90
Fold Enrichment

C

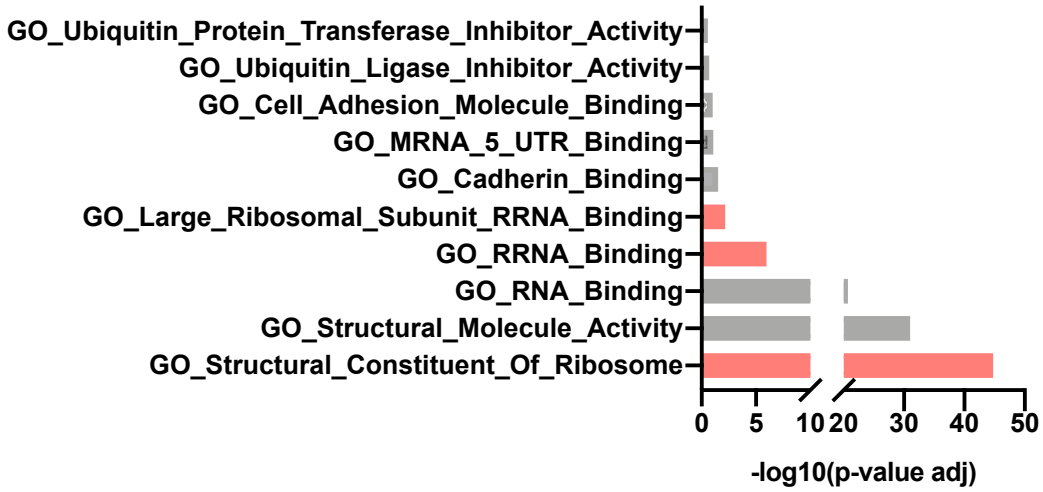

GO_Ubiquitin_Protein_Transferase_Inhibitor_Activity
GO_Ubiquitin_Ligase_Inhibitor_Activity
GO_Cell_Adhesion_Molecule_Binding
GO_MRNA_5_UTR_Binding
GO_Cadherin_Binding
GO_Large_Ribosomal_Subunit_RRNA_Binding
GO_RRNA_Binding
GO_RNA_Binding
GO_Structural_Molecule_Activity
GO_Structural_Constituent_Of_Ribosome
0
5
10
20
30
40
50
-log10(p-value adj)

Figure 4.11: DESeq2 analysis for *Kat2a* WT vs *Kat2a* NULL 4months comparison.
(A) Volcano plot representing differentially expressed genes obtained from comparing *Kat2a* WT vs *Kat2a* NULL cells (4 months). The genes present above the grey line (p= 0.05) are significant and highlighted in red, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-adj <0.05) based on Fisher's exact test with Bonferroni correction, where pink highlights genes associated with translation/ribosome biogenesis, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

To study the common set of genes which are differentially expressed upon *Kat2a* loss at both 2 months and 4 months post transplantation, an intersection was performed between the downregulated genes at both the time points, which yielded 90 genes in common (Fig 4.12A). These 90 genes were fed into Panther gene ontology enrichment analysis, which confirmed the downregulation of translation and ribosomal biogenesis pathway associated genes, in line with the observations in previous comparisons (Fig 4.12B). This was further confirmed using MSigDB overlap, where downregulation of ribosomal structural genes was observed (Fig 4.12C). These observations strengthened the finding that the translation/ribosomal structural constituent genes, mainly *Rpl* and *Rps* gene family, are downregulated upon loss of *Kat2a*. The downregulation of these genes is putatively important for early *RUNX1-RUNX1T1(9a)* leukaemia transformation as well as during leukaemia maintenance.

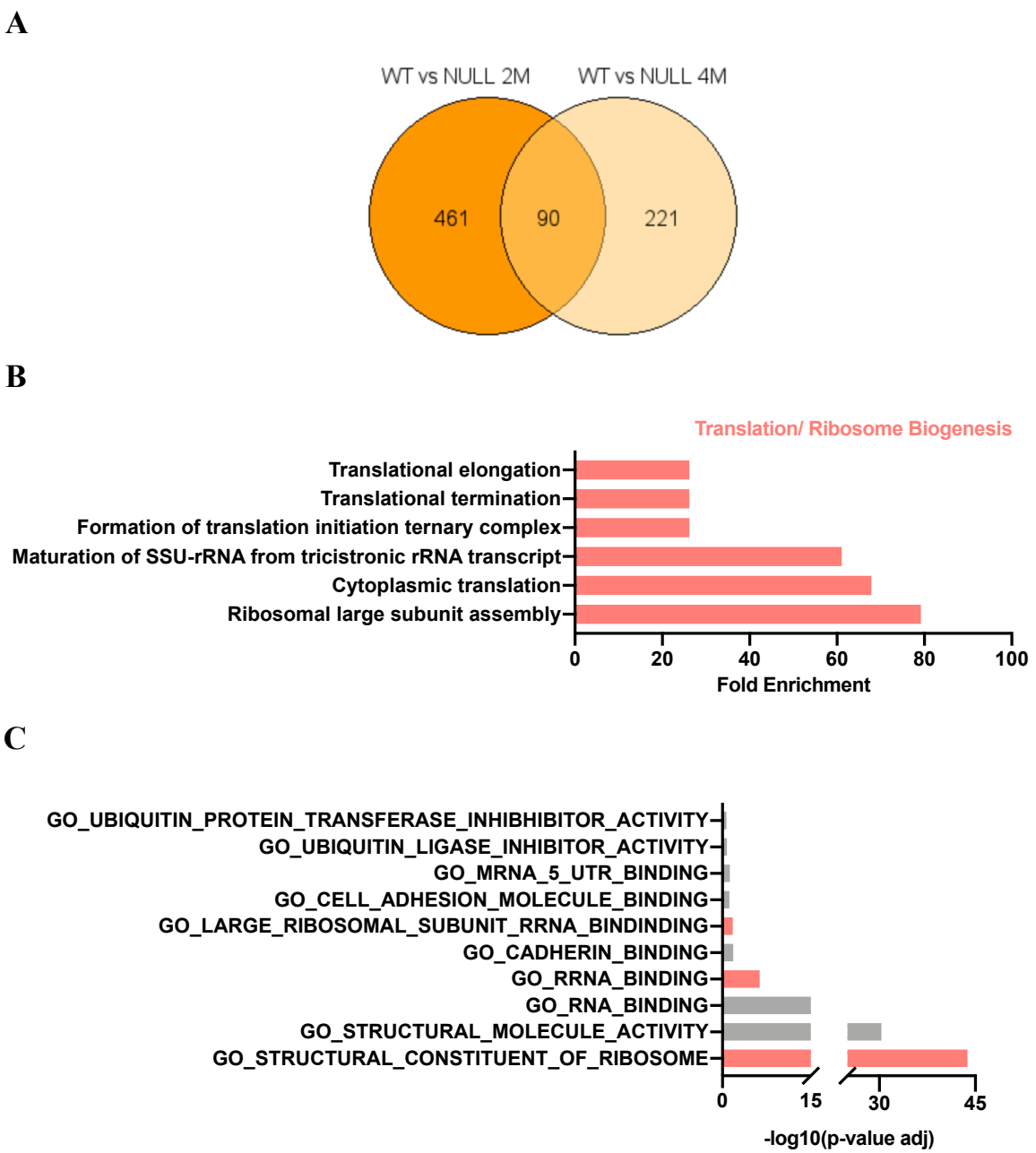


Figure 4.12: DESeq2 analysis for common set of genes downregulated in *Kat2a* NULL with respect to *Kat2a* WT at respective time points.

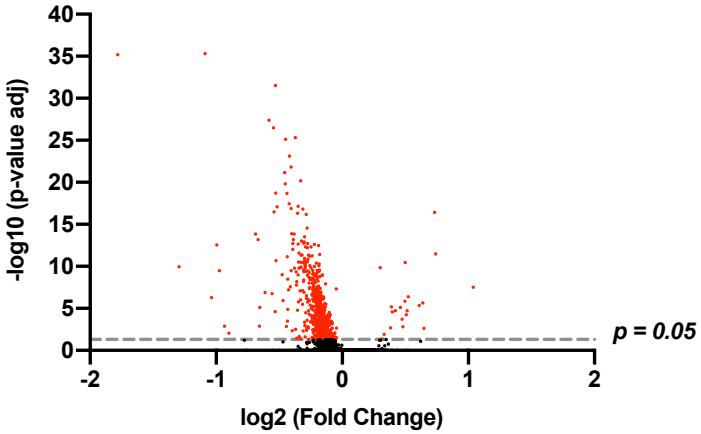
(A) Venn diagram highlighting 90 common genes downregulated in *Kat2a* NULL compared to *Kat2a* WT at the respective time points, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-adj<0.05) based on Fisher's exact test with Bonferroni correction where pink highlights genes associated with translation/ribosome biogenesis, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

4.6 Time series progression of *Kat2a* WT and *Kat2a* NULL suggest an early metabolic configuration, which may be accelerated by *Kat2a* loss

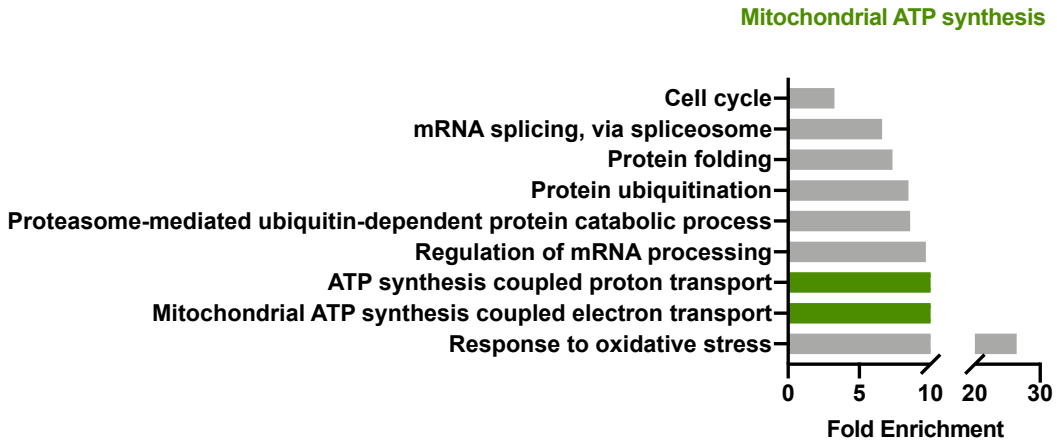
After looking at the gene expression signatures attributed to loss of *Kat2a*, I wanted to study if there are any differences in time series progression of *Kat2a* WT and *Kat2a* NULL. For this, I started with *Kat2a* WT and performed differential expression analysis using DESeq2 at 4 months compared to 2 months post transplantation. There were in total 621 differentially expressed genes, out of which 600 were downregulated, indicating loss of gene expression during *RUNX1-RUNX1T1(9a)* pre-leukaemia progression in *Kat2a* WT cells (Fig 4.13A). Panther gene ontology analysis suggested an enrichment in mitochondrial ATP synthesis pathway genes indicating an early metabolic reconfiguration during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation which may be accelerated upon loss of *Kat2a* (Fig 4.13B). This was consistent with MSigDB overlap approach where pathways associated with oxidoreductase activity, ATPase activity, and proton transmembrane activity were enriched (Fig 4.13C). Other enriched gene expression programmes were found to be associated with regulation of cell cycle genes and post-transcriptional regulatory processes like alternative splicing, along with protein ubiquitination (Fig 4.13B and C).

After comparing the gene expression programmes impacted during *RUNX1-RUNX1T1(9a)* pre-leukaemia progression in *Kat2a* WT cells, similar analysis was done for *Kat2a* NULL cells. Differential expression analysis using DESeq2 yielded 1325 downregulated genes and 79 upregulated genes (Fig 4.14A). Based on Panther enrichment analysis, the downregulated genes were broadly enriched for categories related to regulation of gene expression, including alternative splicing, cis-splicing along with cell cycle regulation, protein modifications, and chromatin organization (Fig 4.14B). This was consistent with the insights from MSigDB overlap, where genes related to RNA binding, transcription factor binding, and protein modifications were enriched (Fig 4.14C). These observations altogether suggested that unlike in *Kat2a* WT cells which mostly have an impact on genes associated with mitochondrial alterations during pre-leukaemia development, *Kat2a* NULL cells follow pre-leukaemia hierarchy by impacting post-transcriptional and epigenetic regulatory mechanisms during the process of pre-leukaemia maintenance.

A


40
35
30
25
20
15
10
5
0
-log10 (p-value adj)
p = 0.05
-2
-1
0
1
2
log2 (Fold Change)

B


Mitochondrial ATP synthesis
Cell cycle
mRNA splicing, via spliceosome
Protein folding
Protein ubiquitination
Proteasome-mediated ubiquitin-dependent protein catabolic process
Regulation of mRNA processing
ATP synthesis coupled proton transport
Mitochondrial ATP synthesis coupled electron transport
Response to oxidative stress
0
5
10
20
30
Fold Enrichment

C

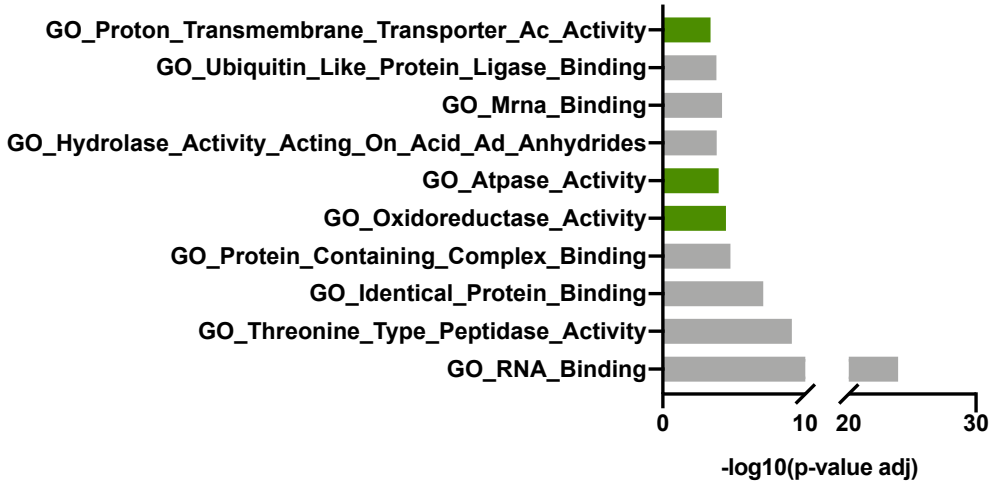
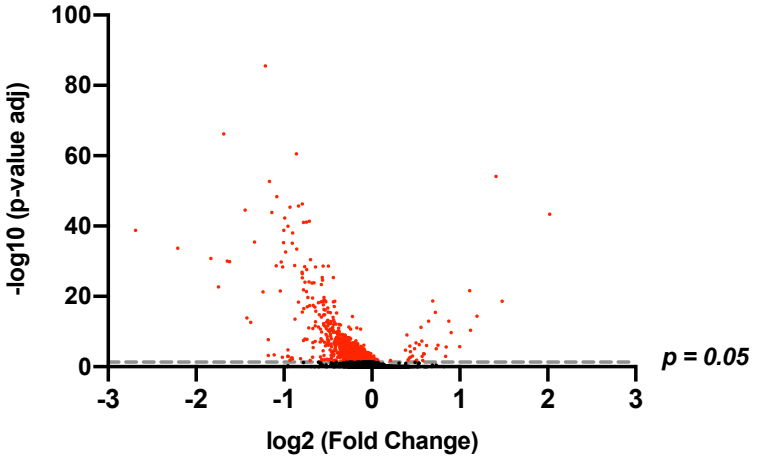

GO_Proton_Transmembrane_Transporter_Ac_Activity
GO_Ubiquitin_Like_Protein_Ligase_Binding
GO_Mrna_Binding
GO_Hydrolase_Activity_Acting_On_Acid_Ad_Anhydrides
GO_Atpase_Activity
GO_Oxidoreductase_Activity
GO_Protein_Containing_Complex_Binding
GO_Identical_Protein_Binding
GO_Threonine_Type_Peptidase_Activity
GO_RNA_Binding
0
10
20
30
-log10(p-value adj)

Figure 4.13: DESeq2 analysis for *Kat2a* WT 2 months vs 4months comparison.
(A) Volcano plot representing differentially expressed genes obtained from comparing *Kat2a* WT 2 months vs 4 months. The genes present above the grey line (p= 0.05) are significant and highlighted in red, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-value adj<0.05) based on Fisher's exact test with Bonferroni correction, where green highlights mitochondrial ATP synthesis associated genes, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

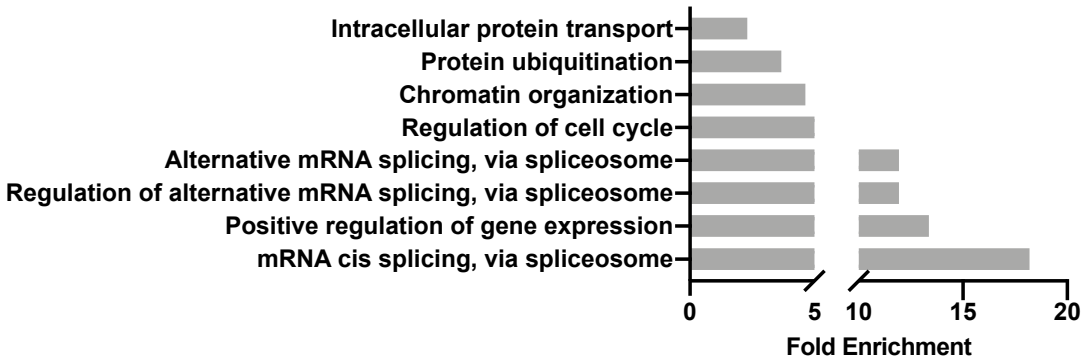
4.7 Mitochondrial ATP synthesis pathway was associated with pre-leukaemia transformation of *Kat2a* WT cells

Having studied gene expression programmes associated with time series progression individually for both *Kat2a* WT and *Kat2a* NULL cells transformed with *RUNX1-RUNX1T1(9a)*, I wanted to understand the gene expression programmes during pre-leukaemia transformation between *Kat2a* NULL at 2 months and *Kat2a* WT at 4 months post transplantation. This comparison would allow to gain insight into the underlying differences between their pre-leukaemia progression hierarchy. I performed differential expression analysis using DESeq2 which gave 493 downregulated and 123 upregulated genes in *Kat2a* NULL cells at 2 months (Fig 4.15A). The downregulated genes were significantly enriched in mitochondrial ATP synthesis and translation/ribosome biogenesis associated gene expression programmes (Fig 4.15B). This was in agreement with MSigDB overlap (Fig 4.15C). Overall, these observations were in-line with the previous observations suggesting an early metabolic reconfiguration which may be accelerated upon loss of *Kat2a* along with translation/ribosome biogenesis associated genes having a role in pre-leukaemia maintenance.

A


100
80
60
40
20
0
-log10 (p-value adj)
-3
-2
-1
0
1
2
3
log2 (Fold Change)
p = 0.05

B


Intracellular protein transport
Protein ubiquitination
Chromatin organization
Regulation of cell cycle
Alternative mRNA splicing, via spliceosome
Regulation of alternative mRNA splicing, via spliceosome
Positive regulation of gene expression
mRNA cis splicing, via spliceosome
0
5
10
15
20
Fold Enrichment

C

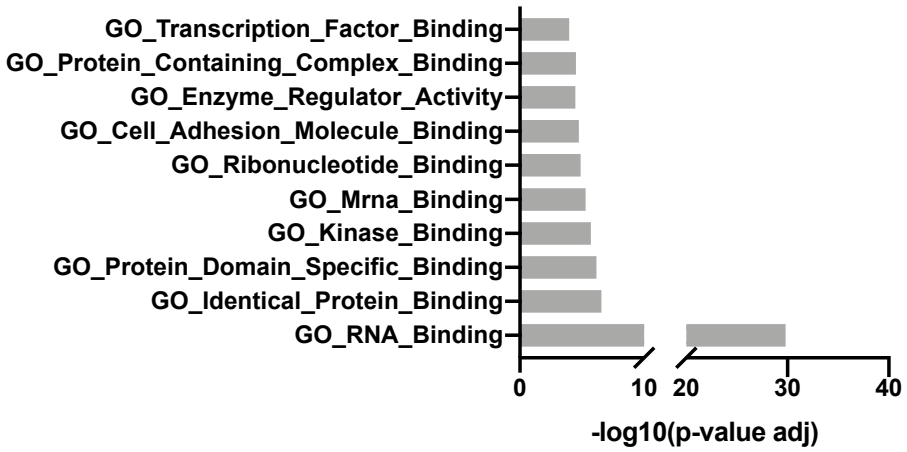
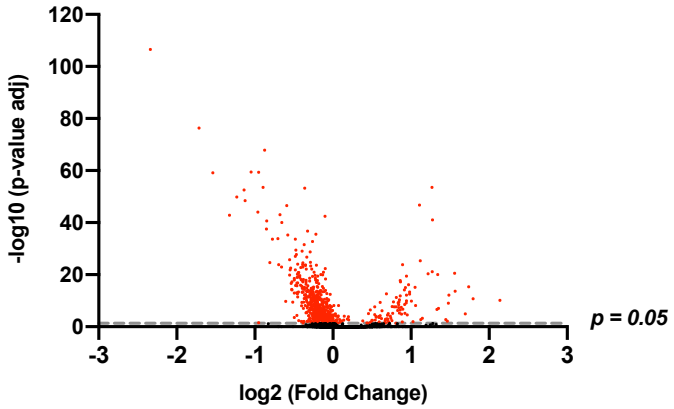

GO_Transcription_Factor_Binding
GO_Protein_Containing_Complex_Binding
GO_Enzyme_Regulator_Activity
GO_Cell_Adhesion_Molecule_Binding
GO_Ribonucleotide_Binding
GO_Mrna_Binding
GO_Kinase_Binding
GO_Protein_Domain_Specific_Binding
GO_Identical_Protein_Binding
GO_RNA_Binding
0
10
20
30
40
-log10(p-value adj)

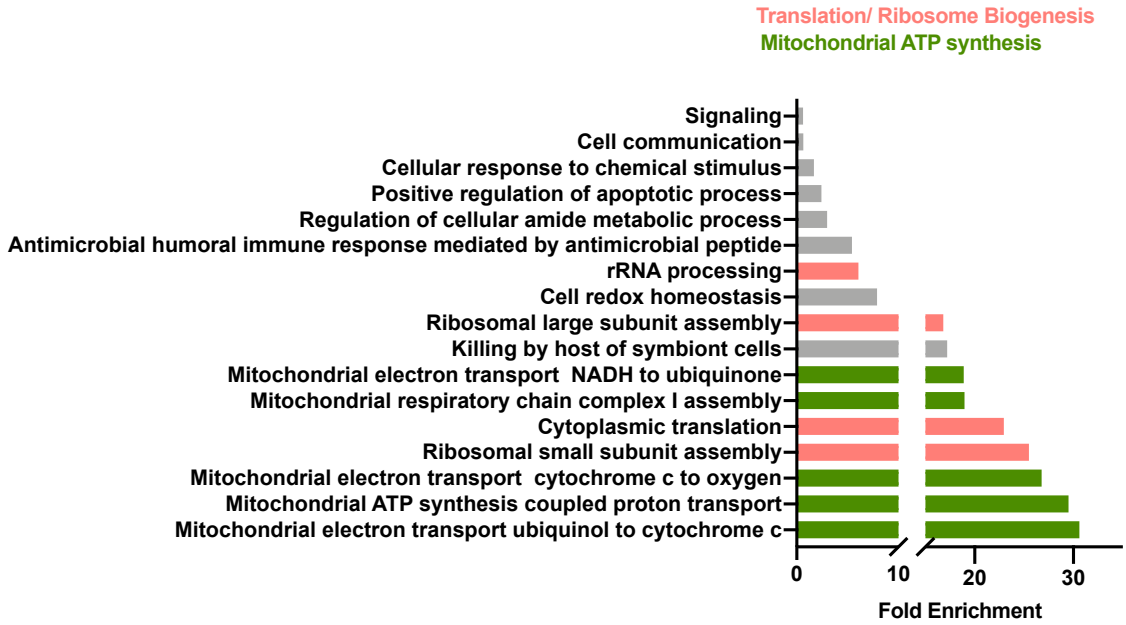
Figure 4.14: **DESeq2 analysis for *Kat2a* NULL 2 months vs 4months comparison.**
(A) Volcano plot representing differentially expressed genes obtained from comparing *Kat2a* NULL 2 months vs 4 months. The genes present above the grey line (p= 0.05) are significant and highlighted in red, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-value adj <0.05*) based on Fisher's exact t-test with Bonferroni correction, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

Finally, I studied the common gene expression pathways expressed differentially in *Kat2a* NULL at 2 months or *Kat2a* WT at 4 months, compared to *Kat2a* WT at 2 months. This comparison would develop an understanding of the gene expression programmes specifically associated with pre-leukaemia transformation of *Kat2a* WT cells at 2 months. For this, I performed an intersection between the list of downregulated genes for the two comparisons (Fig 4.16A). The intersection resulted in a list of 262 genes which were majorly enriched in mitochondrial ATP synthesis processes, highlighting the regulatory role of these processes during pre-leukaemia progression in both *Kat2a* WT at 4 months and *Kat2a* NULL at 2 months (Fig 4.16B and C). The observation of presence of early metabolic reconfiguration during the process of pre-leukaemia transformation was consistent with previous comparisons. However, the fact that this comparison did not capture genes associated with translation/ribosome biogenesis, highlighted the specificity of transcriptional instability in these programmes upon loss of *Kat2a*.

A


-log10 (p-value adj)
120
100
80
60
40
20
0
-3
-2
-1
0
1
2
3
log2 (Fold Change)
p = 0.05

B


Translation/ Ribosome Biogenesis
Mitochondrial ATP synthesis
Signaling
Cell communication
Cellular response to chemical stimulus
Positive regulation of apoptotic process
Regulation of cellular amide metabolic process
Antimicrobial humoral immune response mediated by antimicrobial peptide
rRNA processing
Cell redox homeostasis
Ribosomal large subunit assembly
Killing by host of symbiont cells
Mitochondrial electron transport NADH to ubiquinone
Mitochondrial respiratory chain complex I assembly
Cytoplasmic translation
Ribosomal small subunit assembly
Mitochondrial electron transport cytochrome c to oxygen
Mitochondrial ATP synthesis coupled proton transport
Mitochondrial electron transport ubiquinol to cytochrome c
0
10
20
30
Fold Enrichment

C

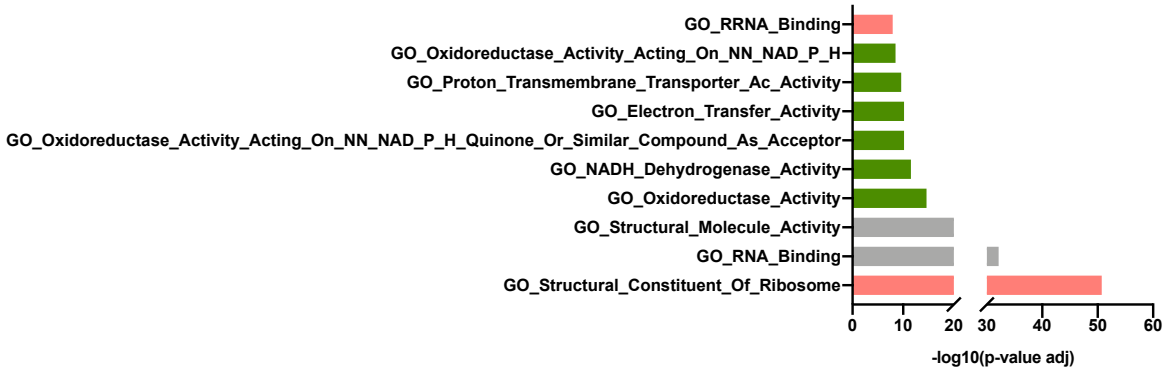

GO_RRNA_Binding
GO_Oxidoreductase_Activity_Acting_On_NN_NAD_P_H
GO_Proton_Transmembrane_Transporter_Ac_Activity
GO_Electron_Transfer_Activity
GO_Oxidoreductase_Activity_Acting_On_NN_NAD_P_H_Quinone_Or_Similar_Compound_As_Acceptor
GO_NADH_Dehydrogenase_Activity
GO_Oxidoreductase_Activity
GO_Structural_Molecule_Activity
GO_RNA_Binding
GO_Structural_Constituent_Of_Ribosome
0
10
20
30
40
50
60
-log10(p-value adj)

Figure 4.15: DESeq2 analysis for *Kat2a* WT 4 months vs *Kat2a* NULL 2 months comparison. **(A)** Volcano plot representing differentially expressed genes obtained from comparing *Kat2a* WT 4 months vs *Kat2a* NULL cells 2 months. The genes present above the grey line (p= 0.05) are significant and highlighted in red, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-adj<0.05) based on Fisher's exact test with Bonferroni correction, where pink highlights genes associated with translation/ribosome biogenesis, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

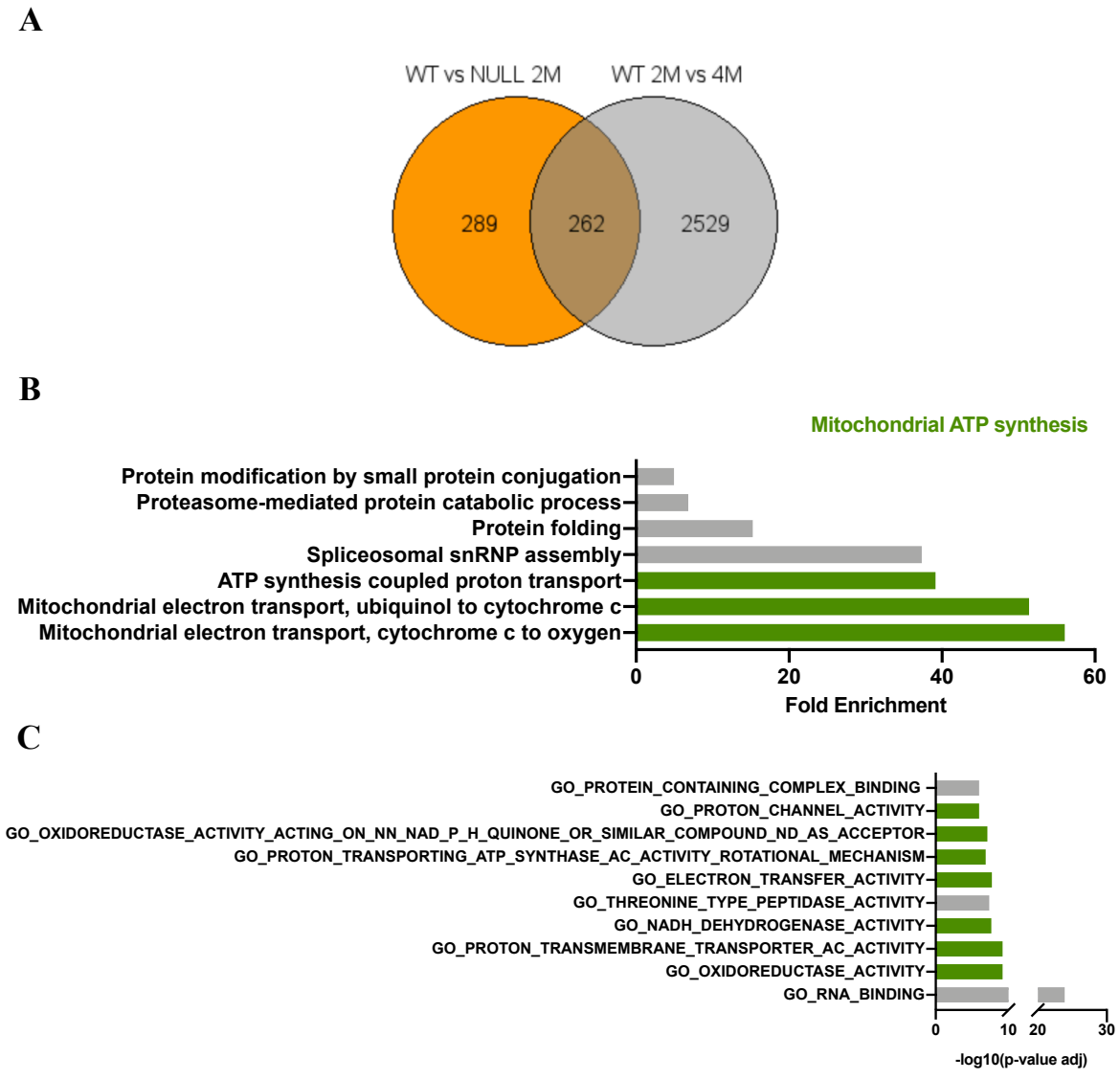


Figure 4.16: DESeq2 analysis for common set of genes downregulated in *Kat2a* NULL vs *Kat2a* WT at 2 months with *Kat2a* WT 2 months vs 4 months.
(A) Venn diagram highlighting 262 common genes downregulated in *Kat2a* NULL vs *Kat2a* WT at 2 months with *Kat2a* WT 2 months vs 4 months, **(B)** Gene Ontology analysis using PANTHER v14 representing significantly enriched categories (p-adj <0.05) based on Fisher's exact test with Bonferroni correction where green highlights genes associated with Mitochondrial ATP synthesis, **(C)** MSigDB v7.1 overlap with C2 dataset highlighting top 10 gene sets with FDR less than 0.05.

In this chapter, I discussed the single-cell RNA sequencing of early-stage *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT and *Kat2a* NULL pre-leukaemia using 10X genomics technology. The samples for scRNA-seq were prepared from frozen bone marrow samples collected from both *Kat2a* WT and *Kat2a* NULL pre-leukaemia animals 2 months and 4 months post transplantation. In total, 1739 cells were subjected to sequencing with 174,770 mean reads per cell. The cell capture efficiency was ~21.73%, which is low compared to the expected 50-65% (Ziegenhain, Vieth, Parekh, Reinius, *et al.*, 2017). This could be attributed to poor sample quality post sorting or limitations associated with library preparation. With recent advancements in 10X technology with v3 being the latest version of the technology, the cell capture efficiency has been improved to ~70%. I obtained 8-10 million reads/sample, much higher than the recommended 1 million (Ziegenhain, Vieth, Parekh and Heyn, 2017), suggesting that the sequencing was performed to a reasonable level of saturation. For individual sample, the number of reads mapped to the genome had fallen in the range of 94-95%, highlighting that it was wise to select mm10 as reference genome. More than 90% of the reads in individual sample were confidently mapped to the genome confirming a good separation from background technical noise. Nearly 70% of the reads in each sample corresponded to transcriptomic region validating the 10X transcriptomics methodology however, it's worth appreciating that the transcripts captured would map to the 3' end of the due to oligodT selection step. Taking the number of detected genes per cell as a measure of sensitivity, I found that individual sample from this study had a median number of genes falling in the range of 1000-2000 genes. One of the major reasons for such low number of detected genes is the fact that the cells utilized for sequencing were cryopreserved for several months and weren't transcriptionally active. This number was of course relatively much lower compared to rest of the single-cell transcriptomics technology where Drop-seq and MARS-seq are capable of

detecting ~5000 median genes per cell whereas Smart-seq/C1 is capable of capturing ~7500 median genes per cell. On the other hand, recently developed Smart-seq2 is capable of detecting up to ~9100 median genes per cell, becoming the most sensitive technique amongst all (Ziegenhain, Vieth, Parekh and Heyn, 2017). However, after comparing the total number of genes detected across many cells, all of these methods including 10X genomics technology were capable of capturing majority of the genes (~13000 genes) where method-specific genes were detected in very few cells (87% of genes occur in one or two cells) (Ziegenhain, Vieth, Parekh and Heyn, 2017). These indications overall highlighted that although 10X genomics doesn't seem comparable to other methods in terms of sensitivity, however, given that the method is capable of detecting majority of the genes, it's highly unlikely that the impact of low median gene count on conclusions drawn from scRNA-seq data will be significant.

It is worth noting that the *Kat2a* NULL sample collected 2 months post transplantation, unlike other samples, did not have a clear distinction between the cells detected and background noise with merely 2753 median UMI counts per cell as opposed to more than 6000 UMI counts per cell in other samples. I hypothesize that this not-so-clear distinction could be due to higher percentage of terminally differentiated cells in the sample, which may not acclimatize to the freezing-thawing cycle very well, and hence, the sample quality may not be comparable to the rest of the samples. Again, with recent developments and an improved CellRanger pipeline, these challenges associated with clear distinction between cells and noisy background along with dropout rates have been addressed remarkably.

Due to inherent noisiness of scRNA-seq data (Hwang, Lee and Bang, 2018), owing to various technical and biological factors, I first performed a clean-up of the scRNA-seq data obtained above (Butler *et al.*, 2018). For this, I started with imposing two different types of cut-offs, first, by keeping the minimum number of genes to 500, in-line with the previous lab study (Domingues *et al.*, 2020), where each cell having gene expression values for less than 500 genes was considered a poor quality cell. Secondly, a cut-off was imposed on number of cells, where each gene was considered to have a significant expression in the given biological system if a minimum number of cells showed expression for it. To begin with, three such different cut-offs (2%, 3% and 4%) for minimum number of cells were imposed for downstream analysis. Then, I excluded the cells with compromised viability as highlighted by high mitochondrial

gene expression, which is indicative of apoptotic or necrotic cells. Further, potential doublets or multiplets were filtered out based on abnormal gene counts represented by a barcode representative of gene expression data from more than a single cell. Post-filtering, normalization was performed to account for any cell-specific bias and zero-inflated counts due to reasons such as dropout or transient gene expression (Hwang, Lee and Bang, 2018). The normalization was performed using log-transformation method and a scaling factor which is estimated by standardizing across the given cells, assuming that most genes are not differentially expressed. Although, log-normalization, so-far has been the most reliable method for researchers handling scRNA-seq data, however, it's worth considering that certain factors including differences in cell lysis, reverse transcription efficiency, and stochastic molecular sampling during sequencing also contribute significantly towards technical bias and may potentially impact downstream analysis (Hafemeister and Satija, 2019). The data variability due to confounding technical challenges associated with PCR has somewhat been resolved by the incorporation of unique molecular identifiers (UMI) in scRNA-seq (Islam *et al.*, 2014b). However, the potential for combining within-sample and between-sample normalization methods still largely remains unexplored and an active area of research that will require rigorous testing.

Post-filtering and normalization of scRNA-seq data, I focussed on highly variable genes for further downstream analysis. I used a function implemented in Seurat v2.4 which calculates the average expression and dispersion for each gene, places these genes into bins, and then calculates a z-score for dispersion within each bin. This function gave an output of ~2000 genes at the individual cut-off imposed at the beginning of filtering process. The 4% cut-off yielded only 1,184 highly variable genes, indicating that this cut-off may not be able capture rare-cell populations compared to the 2% and 3% cut-off, which yielded 2,063 and 1,844 highly variable genes. These highly variable genes thus obtained were utilised for performing linear dimensionality reduction analysis using Principal Component Analysis (PCA). PCA helped in defining the significant principal components contributing towards gene expression variability, thus identifying the true dimensionality of the dataset.

Based on the significant principal components identified using PCA, a graph-based clustering approach was followed. The cells were further clustered using a modularity optimization

technique, the Louvain algorithm (default) which iteratively grouped cells together, with the goal of optimizing the standard modularity function (Vincent D Blondel *et al.*, 2008). This is an improved approach compared to conventional approaches which follow Euclidean norm or cosine distance, as these are not effective enough for high-dimensional data with few objects. (Beyer *et al.*, 1998).

In order to lay over the clusters obtained in a low dimensional space, t-Distributed Stochastic Neighbour Embedding (t-SNE) (Van Der Maaten and Hinton, 2008) was used which placed the cells with similar local neighbourhoods in high-dimensional space together in low-dimensional space. The reduced dimensional data thus obtained had 5 clusters with less well-defined boundaries in case of the 2% cut-off, whereas well-segregated clusters were obtained in case of the 3% (6 clusters) and 4% cut-offs (8 clusters). Combining information from highly variable genes and obtained clusters, I utilised 3% cut-off for downstream analysis in order to obtain high confidence rare cell populations without losing informative genes. Overall, the pre-processing of scRNA-seq data obtained from early-stage *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT and *Kat2a* NULL pre-leukaemia animals yielded a gene expression matrix of 1675 cells with 1575 median genes having 5939 median UMI counts per cell.

After performing the clean-up steps on raw scRNA-seq data, the transcriptional programmes associated with *Kat2a* loss were studied and correlated with pre-leukaemia progression. Differential expression analysis was performed using DESeq2 (Love, Huber and Anders, 2014). The global comparison of *Kat2a* NULL with respect to *Kat2a* WT, revealed that loss of *Kat2a* leads to an overall downregulation of gene expression, with 811 genes (p-adj <0.05) being downregulated compared to 70 genes being upregulated (p-adj <0.05). Interestingly, the global comparison also suggested that there were only 82 genes out of 881 having more than 20% change in gene expression upon *Kat2a* loss. These observations were compatible with the acetyltransferase activity of *Kat2a*, playing an important role in promoting transcriptional stability (Hebbes, Thorne and Crane-Robinson, 1988)(Turner BM, 1993). Altogether, these inferences from literature are held in conjunction with the observation that loss of *Kat2a* promotes transcriptional instability during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation.

After looking at the global comparison, gene expression changes were studied upon *Kat2a* loss by comparing *Kat2a* NULL cells to *Kat2a* WT cells at individual time points. A comparison between 2 months samples would indicate *Kat2a* associated transcriptional changes during pre-leukaemia initiation whereas comparison between 4 months samples would indicate transcriptional changes associated with *Kat2a* loss during pre-leukaemia maintenance. There were 551 downregulated and 48 upregulated genes upon loss of *Kat2a* at 2 months and 118 downregulated and 39 upregulated genes upon loss of *Kat2a* at 4 months post transplantation. These analyses highlighted loss of *Kat2a* may potentially impact gene expression during early stages of pre-leukaemia transformation. I also compared the transcriptional changes in *Kat2a* WT occurring between 2 months and 4 months and *Kat2a* NULL occurring between 2 months and 4 months separately. *Kat2a* WT cells showed a downregulation of 600 genes at 4 months compared to 2 months, whereas 1,325 genes were downregulated in *Kat2a* NULL cells at 4 months compared to 2 months. The higher number of downregulated genes in *Kat2a* NULL time-series comparison may indicate that *Kat2a* loss leads to downregulation of specific sets of genes during the process of leukaemia progression. This also suggested that downregulation of gene expression programmes may be associated with pre-leukaemia transformation. This observation may not be surprising given that pre-leukaemic clones contribute towards disease relapse which may be a consequence of reduced transcription of genes (Lee *et al.*, 2006; Majeti *et al.*, 2009).

I then sought to understand the pathways associated with *Kat2a* loss at a global scale, Translation/ribosome biogenesis and mitochondrial ATP synthesis pathways were found to be the two major enriched categories. These observations were compatible with a previous study conducted in our lab on *MLL-AF9* leukaemia, where loss of *Kat2a* downregulated gene expression pathways associated with translation/ribosome biogenesis and mitochondrial ATP synthesis (Domingues *et al.*, 2020). These findings altogether suggested that loss of *Kat2a* may exert its effect through general pathways rather than leukaemia specific ones. Similar analysis was performed for genes downregulated at individual time-points. There were 599 significantly differentially expressed genes in *Kat2a* NULL compared to *Kat2a* WT at 2 months post transplantation, out of which 551 genes were downregulated. Gene ontology analysis indicated enrichment of similar gene expression programmes as global comparisons, namely translation/ribosomal biogenesis and mitochondrial ATP synthesis, highlighting their

importance in pre-leukaemia transformation. Similarly, I looked at the pathways enriched in *Kat2a* NULL at 4 months post transplantation. There were 157 genes in total which were differentially expressed at 4 months, a number much lower than the 599 differentially expressed genes at 2 months, indicating that loss of *Kat2a* leads to transcriptional changes at early stages of *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation. The 118 downregulated genes out of the 157 genes were subjected to gene ontology enrichment analysis, which again suggested overrepresentation of genes related to translation and ribosome biogenesis. Unlike previous comparisons performed globally and at 2 months, no other categories were enriched, perhaps indicating that the missing pathways were specifically required at early stages of transformation whereas translation/ribosome biogenesis were continually required for pre-leukaemia maintenance. However, this may also indicate that the time point at 4 months post transplantation may be too late to capture alterations in the pathways which were enriched globally and at 2 months. Further, 90 genes were found to be differentially expressed upon *Kat2a* loss at both 2 months and 4 months post transplantation. These 90 genes were also enriched in translation and ribosomal biogenesis pathway associated genes, in line with the observations in previous comparisons.

To understand the time series progression of *Kat2a* WT cells and *Kat2a* NULL cells separately, differential expression analysis was performed comparing the 4 months time point to the 2 months time for each genotype separately. In total, there were 621 differentially expressed genes in *Kat2a* WT cells at 4 months, out of which 600 were downregulated indicating loss of gene expression during *RUNX1-RUNX1T1(9a)* leukaemia progression in *Kat2a* WT cells undergoing pre-leukaemia transformation. The gene ontology analysis suggested an enrichment in mitochondrial ATP synthesis pathway genes, highlighting the presence of metabolic reconfiguration during early stages of pre-leukaemia transformation. These metabolic changes may be key to pre-leukaemia progression and are accelerated in the absence of *Kat2a.* On a similar note, another analysis was performed for *Kat2a* NULL cells comparing 4 months to 2 months post transplantation. Differential expression analysis yielded a total of 1325 downregulated genes and 79 upregulated genes which were broadly enriched for categories related to regulation of gene expression, including alternative splicing, cis-splicing, along with cell cycle regulation, protein modifications, and chromatin organization. The analysis highlights that loss of *Kat2a* impacts translation/ribosomal biogenesis and

mitochondrial ATP synthesis during early stages of pre-leukaemia transformation and potentially needs these processes during pre-leukaemia maintenance. However, once the *RUNX1-RUNX1T1(9a)* pre-leukaemia is established, the disease progression impacts transcriptional programmes associated with cell cycle progression, post-transcriptional modifications, along with epigenetic instability in the form of chromatin reorganization.

Altogether, the analysis presented in this chapter highlights that enrichment of mitochondrial bioenergetics indicate an early metabolic reconfiguration during pre-leukaemia transformation, which may be accelerated upon loss of *Kat2a*. However, the impairment in ribosomal biogenesis and translation associated genes is a consequence of loss of *Kat2a* which maybe continually required for pre-leukaemia maintenance. Based on these insights obtained from scRNA-seq, I sought to understand their mechanistic contribution by inhibition of these pathways individually during the process of *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation which is discussed in the next chapter.

5 Mechanistic investigation of *Kat2a* associated transcriptional programmes in *RUNX1-RUNX1T1(9a)* and *Idh1*R132H pre-leukaemia

In the previous chapter, I identified the prominent gene expression pathways that show transcriptional instability upon loss of *Kat2a* during *RUNX1-RUNX1T1(9a)* pre-leukaemia, including mitochondrial bioenergetics and ribosomal biosynthetic programmes. The downregulation of mitochondrial ATP bioenergetics as well as mitochondrial translation reflect two different aspects of metabolic reprogramming during early pre-leukaemia transformation as a consequence of *Kat2a* loss. The consistent attenuation in ribosomal biogenesis machinery at both time points, however, indicated that reduced protein synthesis may impact *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation *in vivo* upon loss of *Kat2a.* In order to validate these observations experimentally *in vitro*, I made use of *RUNX1-RUNX1T1(9a)* and *Idh1*R132H models of pre-leukaemia to assess the mitochondrial activity and protein synthesis. In parallel, I conducted similar experiments on the *MLL-AF9* model of leukaemia, which represents a strong oncogenic model in contrast to the other two pre-leukaemia models. As discussed in Chapter-1, *MLL-AF9* transformed cells could aid in maintenance of the leukaemia by immortalization of leukaemogenic cell populations, which is dependent on constitutive *MLL-AF9* expression (Somervaille and Cleary, 2006) (Horton *et al.*, 2013). Previous studies conducted in our lab suggested a depletion of *MLL-AF9* leukaemia stem-like cells upon loss of *Kat2a* via promotion of transcriptional instability in mitochondrial bioenergetics and ribosomal biosynthetic programmes (Domingues *et al.*, 2020). The observations of immortalization of AML transformed cells in the presence of strong oncogenic effect of *MLL-AF9* fusion protein suggest it is an excellent model of AML maintenance in contrast to the pre-leukaemia models, with similar transcriptional programmes being impacted upon loss of *Kat2a*.

5.1 *Kat2a* loss downregulates mitochondrial activity in *RUNX1-RUNX1T1(9a)* pre-leukaemia

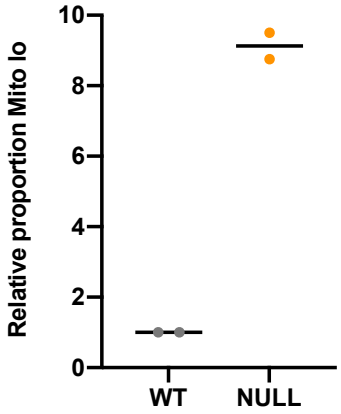
To study the role of mitochondrial activity, I started with collection of BM cells from *RUNX1-RUNX1T1(9a) in vivo* transformed *Kat2a* WT and *Kat2a* NULL animals at 2 months post transplantation. To test whether loss of *Kat2a* impairs mitochondrial functionality, both *Kat2a* WT and *Kat2a* NULL BM cells were subjected to staining with two fluorescent dyes, namely Mitotracker deep red, and Mitostatus TMRE (Methods). Mitotracker deep red contains a mildly thiol-reactive chloromethyl moiety, which passively diffuses across the plasma membrane and accumulates in active mitochondria, hence, indicating active mitochondrial content/mass inside a cell. On the other hand, Mitostatus TMRE measures mitochondrial membrane potential ($\Delta\psi m$) of a cell, which is a measure of oxidative stress, apoptosis, or any other stressful event (Scaduto and Grotyohann, 1999). For example, in cells undergoing apoptosis, pro-apoptotic Bcl-2 family proteins cause mitochondrial outer membrane permeabilization (MOMP), resulting in the release of cytochrome C, and the subsequent activation of caspase-9 and the apoptotic cascade (Gottlieb, Vander Heiden and Thompson, 2000) (Lemasters *et al.*, 1998). This MOMP often correlates with the loss of inner mitochondrial membrane potential ($\Delta\psi m$), which can be detected using $\Delta\psi m$-sensitive dyes. Mitostatus TMRE is one such cationic, lipophilic dyes which gets accumulated within the mitochondria of healthy cells, but not within mitochondria that have lost $\Delta\psi m$ due to induction of apoptosis, or treatment with a mitochondrial uncoupler. To avoid confounding effects due to variability in cell size and mitochondrial number, cells with similar forward and side scatter profile were included in the analysis. I observed that *Kat2a* NULL cells transformed with *RUNX1-RUNX1T1(9a) in vivo* displayed an increase in a population with low active mitochondrial mass (Mito lo) compared to *Kat2a* WT (Fig 5.1A). A representative flow cytometry plot highlighted the shift in the Mitotracker positive population leading to reduced active mitochondrial mass (Fig 5.1B). On the other hand, no changes were observed in terms of mitochondrial potential, as indicated by Mitostatus TMRE, upon loss of *Kat2a* (Fig 5.1B).

To understand the general role of *Kat2a* in leukaemia progression, I also looked at potential mitochondrial alterations in *MLL-AF9 in vitro* transformed *Kat2a* WT and *Kat2a* NULL BM

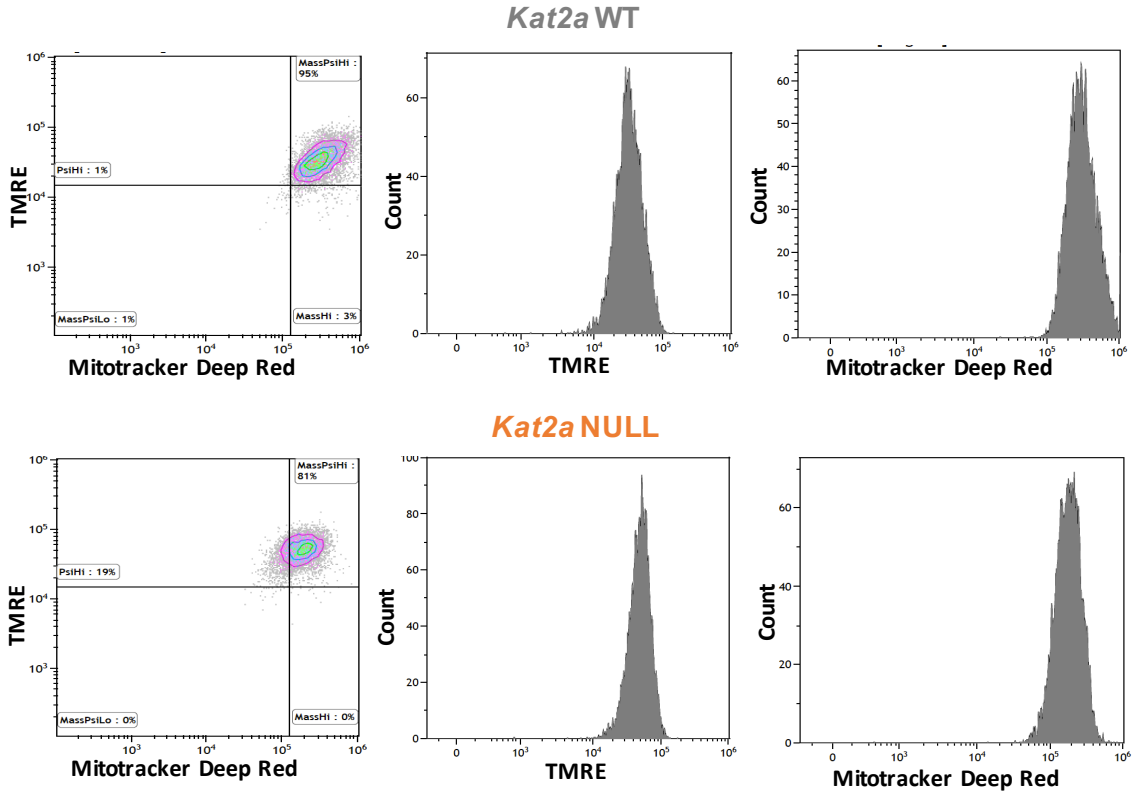
cells. *MLL-AF9* represents a leukaemia maintenance model requiring fewer cooperating mutations, unlike *RUNX1-RUNX1T1(9a)* and *Idh1*R132H. Previous work conducted in our lab suggested that *Kat2a* plays a role in *MLL-AF9* driven transformation by impacting mitochondrial translational machinery and ribosomal biosynthetic programmes (Domingues *et al.*, 2020). To investigate this observation mechanistically, I performed the same experiment with Mitotracker deep red and Mitostatus TMRE in *Kat2a* NULL and *Kat2a* WT cells transformed with *MLL-AF9 in vitro.* As per the findings from single cell RNA sequencing (scRNA-seq) on *MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL cells, I observed an increase in the Mito lo population in *Kat2a* NULL cells compared to *Kat2a* WT cells (Fig 5.1C). This is highlighted by a shift in the Mitotracker positive population, while no changes were observed in the Mitostatus TMRE positive population, in line with the *RUNX1-RUNX1T1(9a)* pre-leukaemia observations (Fig 5.1D).

RUNX1-RUNX1T1(9a)

A

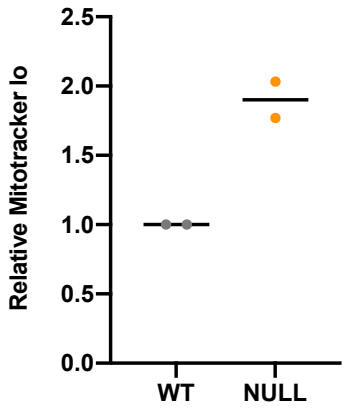


B



MLL-AF9

C


Relative Mitotracker lo
2.5
2.0
1.5
1.0
0.5
0.0
WT
NULL

D

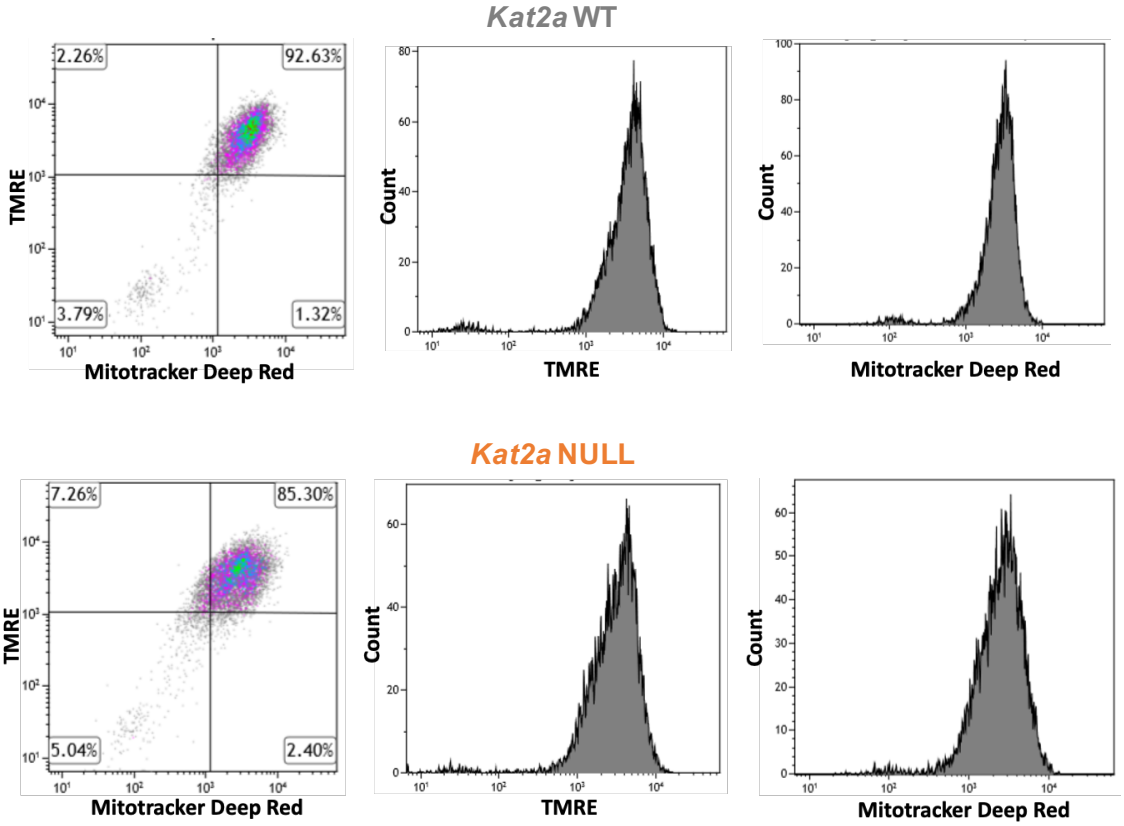

Kat2a WT
2.26%
92.63%
3.79%
1.32%
TMRE
Mitotracker Deep Red
Count
TMRE
Count
Mitotracker Deep Red
Kat2a NULL
7.26%
85.30%
5.04%
2.40%
TMRE
Mitotracker Deep Red
Count
TMRE
Count
Mitotracker Deep Red

Figure 5.1: **Mitochondrial mass and potential analysis for *RUNX1-RUNX1T1(9a)* and *MLL-AF9* transformed cells.**
(A) Relative proportion of *RUNX1-RUNX1T1(9a) in vivo* transformed *Kat2a* NULL BM cells and *Kat2a* WT BM cells that have low mitochondrial mass at 2 months post transplantation (n= 2/genotype), horizontal bars represent mean value and dots indicate individual data points **(B)** Representative flow cytometry dot plot (left) highlighting mitochondrial mass on x-axis and mitochondrial potential on y-axis, along with the respective histograms (right) for *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT BM cells (top) and *Kat2a* NULL BM cells (bottom), **(C)** Relative proportion of *MLL-AF9 in vitro* transformed *Kat2a* NULL BM cells and *Kat2a* WT BM cells that have low mitochondrial mass (n= 2/genotype), **(D)** Representative flow cytometry dot plot (left) highlighting mitochondrial mass on x-axis and mitochondrial potential on y-axis along with respective histograms (right) for *MLL-AF9* transformed *Kat2a* WT BM cells (top) and *Kat2a* NULL BM cells (bottom).

5.2 *MLL-AF9* transformed *Kat2a* WT cells with low mitochondrial mass phenocopy some of the characteristics of *Kat2a* NULL cells

Having confirmed that loss of *Kat2a* reduces mitochondrial content in both *RUNX1-RUNX1T1(9a)* and *MLL-AF9* leukaemia, I wanted to study whether segregating cell populations on the basis of mitochondrial content would allow me to dissect the mechanistic contribution of mitochondrial activity during leukaemia progression. For this, I made use of *MLL-AF9* model of leukaemia, due to its impact on immortalization of leukaemia cells which allows long term maintenance of these transformed cells *in vitro.*

I started with *MLL-AF9 in vitro* transduced *Kat2a* WT and *Kat2a* NULL BM cells, which were maintained in the form of a colony forming assay (CFC assay) in a semi-solid methylcellulose medium supplemented with cytokines, allowing for the maintenance of myeloid progenitor cells. The *MLL-AF9 in vitro* transduced *Kat2a* WT and *Kat2a* NULL BM cells were maintained in the CFC assay for three rounds of plating, until the colonies obtained were enriched in transformed cell populations. These transformed cell populations having either *Kat2a* WT or *Kat2a* NULL genotype were stained with Mitostatus TMRE and Mitotracker deep red fluorescent dye and sorted on the basis of high mitochondrial mass (Mito Hi) and low mitochondrial mass (Mito Lo) (gating strategy described in Methods). The different fractions

of cells obtained post sorting were namely, *Kat2a* WT Mito hi, *Kat2a* WT Mito lo, *Kat2a* NULL Mito hi, and *Kat2a* NULL Mito lo. These cell populations were subjected to CFC assay in order to assess differences in their self-renewal potential.

Globally, the total number of colonies obtained from *Kat2a* NULL fractions were lower than those obtained from *Kat2a* WT fractions, compatible with the observation in *MLL-AF9* model (described in detail in Chapter-6, Figure 6.4H). The cells from Mito lo fractions had lower number of colonies compared to the respective Mito hi fractions for both genotypes, indicating that segregation of cells based on mitochondrial content may not have a genotype-specific effect (Fig 5.2A). However, the number of colonies obtained in *Kat2a* WT Mito lo was comparable to that of *Kat2a* NULL Mito hi and *Kat2a* NULL Mito lo together, suggesting that *Kat2a* NULL as a whole phenocopies *Kat2a* WT Mito lo.

I then looked at the colony composition of individual cell subsets based on criteria defined in a study by Lavau and colleagues (Lavau *et al.*, 1997), where colonies were subdivided into compact, mixed, and dispersed types. As the nomenclature suggests, compact colonies are the ones which are compact and do not show any halo of migrating cells, mixed colonies have a compact centre with a diffuse halo of differentiating progenitors, and dispersed colonies are comprised of large diffuse colonies without a defined centre. Compact colonies are determinant of the presence of immature myeloid cells, whereas diffuse components include differentiated macrophages. Based on these criteria, looking at the different colony types in these cellular fractions, I observed a reduction in compact colonies in *Kat2a* NULL compared to *Kat2a* WT cells, compatible with the earlier observation (Chapter-6). In line with the reduction in total colonies, there was also a significant reduction in compact colonies from Mito hi to Mito lo fractions in both genotypes, indicating a reduction in colony forming potential in the Mito lo population, irrespective of genotype (Fig 5.2B). Once again, the relative proportion of compact colonies from *Kat2a* WT Mito lo was comparable to that of *Kat2a* NULL Mito hi and *Kat2a* NULL Mito lo, suggesting that *Kat2a* NULL as a whole phenocopies *Kat2a* WT Mito lo, a trend compatible with total colonies. Further, I looked at the proportion of mixed and dispersed colonies which is determinant of cellular differentiation in all four populations. In line with previous observations (Chapter-6), mixed colonies seemed specific to *MLL-AF9* transformed *Kat2a* NULL cells and *Kat2a* NULL Mito hi captured most of these. There was a significant

reduction in mixed colony proportion in *Kat2a* NULL Mito lo cells relative to *Kat2a* NULL Mito hi cells (Fig 5.2C). The proportion of mixed colonies in *Kat2a* NULL Mito lo was comparable to the mitochondrial fractions in *Kat2a* WT cells. Dispersed colonies were the lowest represented colonies in all the samples, consistent with the previous observations (Chapter-6), which made it challenging to identify subtle significant differences in the proportion of these colonies in different samples. However, there was a trend towards an increase in the proportion of dispersed colonies in *Kat2a* WT Mito lo, comparable with global *Kat2a* NULL cells, again indicating that *Kat2a* NULL as a whole phenocopies *Kat2a* WT Mito lo (Fig 5.2D). In summary, this experiment strongly suggests that reduction in active mitochondrial mass impacts self-renewal potential of *MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL mouse bone marrow cells. After segregation based on active mitochondrial mass, these cells may lose their colony forming potential as observed in cells with lower mitochondrial mass, where loss of *Kat2a* mimics the replating capacity of *Kat2a* WT Mito lo cells (Fig 5.2E).

After studying the functional impact of loss of *Kat2a* on active mitochondrial mass and how this affects the self-renewal potential of *MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL cells, I aimed to understand the impact of active mitochondrial mass on clonal expansion of *MLL-AF9* transformed cells. For this, *MLL-AF9 in vitro* transformed *Kat2a* WT and *Kat2a* NULL cells were stained with Mitostatus TMRE and Mitotracker deep red, and further index-sorted on the basis of Mitotracker deep red fluorescence (active mitochondrial mass). Individual cell populations were single-cell sorted in a 96-well plate, giving a total of 6 samples- *Kat2a* WT Mito hi, *Kat2a* WT Mito lo, *Kat2a* WT total, *Kat2a* NULL Mito hi, *Kat2a* NULL Mito lo, *Kat2a* NULL total. The index-sorting technology enabled study of clonal expansion capability of individual cells from separate mitochondrial fractions sorted from each genotype (Hayashi *et al.*, 2010). The single-cell clonal cultures were maintained for a week and cell counting was performed for individual wells on a daily basis. The cell counts thus obtained were categorized into 7 intervals, where each interval was coded in a different colour. These counts were then organized in the form of a heatmap hierarchically, so that the cells with increased clonal expansion capacity can be correlated with their colony forming potential. The generated global heatmap suggested a decrease in clonal expansion potential in *Kat2a* NULL cells compared to *Kat2a* WT as seen from the Mito Total, in line with the observation from

colony forming assay (Fig 5.3A). This was also compatible with the insights from colony forming assay discussed in Chapter-6. The cells from Mito lo fractions had reduced clonal expansion capacity compared to the respective Mito hi fractions for both genotypes, indicating that segregation of cells based on mitochondrial content may not have a genotype-specific effect (Fig 5.3B). This was again in accordance with the colony forming assay observations above, where both Mito lo fractions had reduced colony forming potential compared to their respective Mito hi fractions.

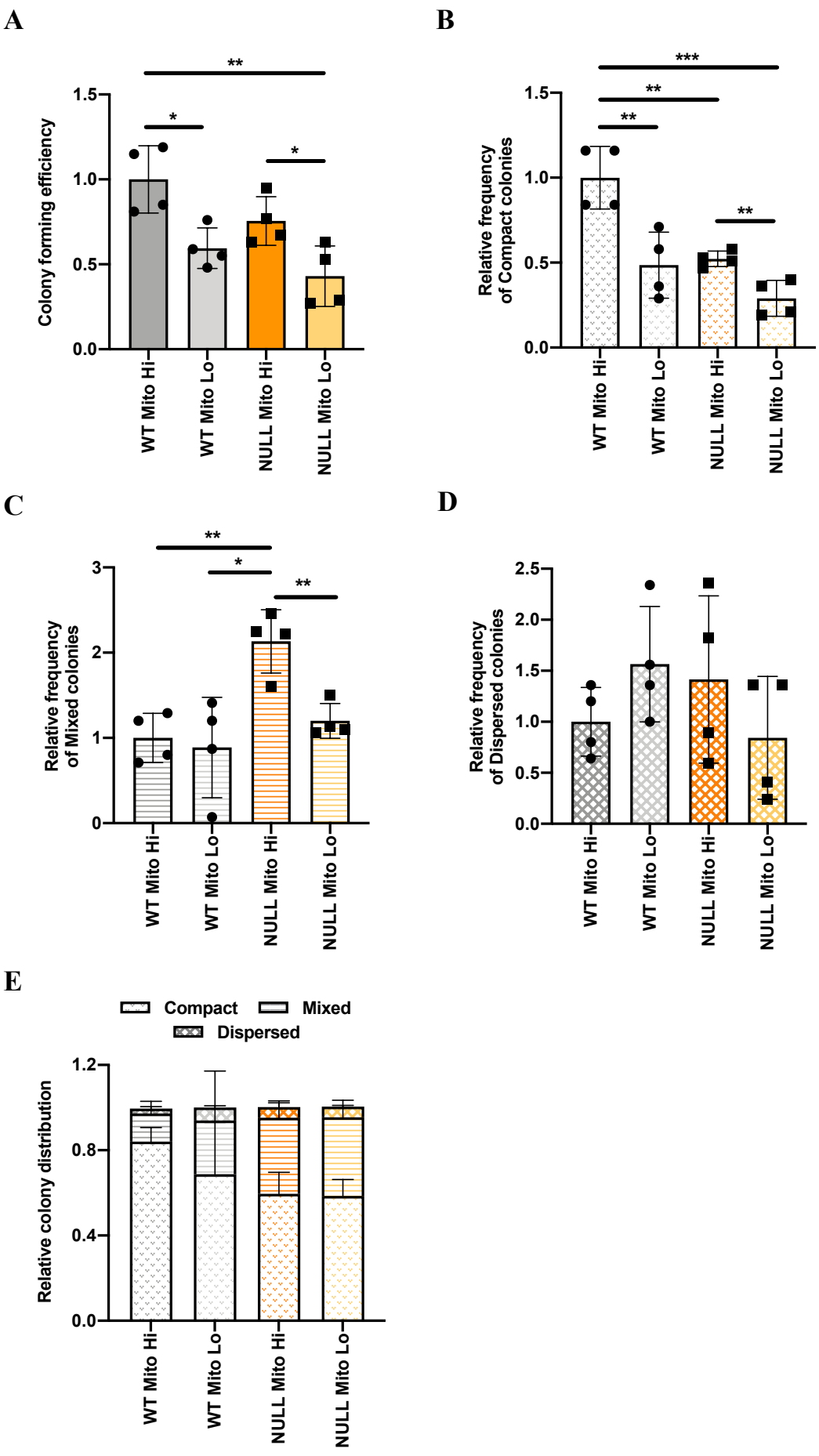

A
Colony forming efficiency
1.5
1.0
0.5
0.0
**
*
*
WT Mito Hi
WT Mito Lo
NULL Mito Hi
NULL Mito Lo
B
Relative frequency of Compact colonies
1.5
1.0
0.5
0.0
***
**
**
**
WT Mito Hi
WT Mito Lo
NULL Mito Hi
NULL Mito Lo
C
Relative frequency of Mixed colonies
3
2
1
0
**
*
**
WT Mito Hi
WT Mito Lo
NULL Mito Hi
NULL Mito Lo
D
Relative frequency of Dispersed colonies
2.5
2.0
1.5
1.0
0.5
0.0
WT Mito Hi
WT Mito Lo
NULL Mito Hi
NULL Mito Lo
E
Compact
Mixed
Dispersed
Relative colony distribution
1.2
0.8
0.4
0.0
WT Mito Hi
WT Mito Lo
NULL Mito Hi
NULL Mito Lo

Figure 5.2: Colony forming assay for *in vitro MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL BM cells sorted on the basis of mitochondrial mass.

(A) Colony forming frequency for *Kat2a* WT mitochondrial mass high (WT Mito Hi), *Kat2a* WT mitochondrial mass low (WT Mito Lo), *Kat2a* NULL mitochondrial mass high (NULL Mito Hi), *Kat2a* NULL mitochondrial mass low (NULL Mito lo) (n=4/sample, mean ± SD, Nested t-test, p= 0.017* WT Mito hi vs lo, p= 0.031* NULL Mito hi vs lo, p= 0.0052** WT Mito hi vs NULL Mito lo), **(B)** Relative frequency of compact colonies (n=4/sample, mean ± SD, Nested t-test, p= 0.0085** WT Mito hi vs lo, p= 0.0068** NULL Mito hi vs lo, p= 0.0024** WT Mito hi vs NULL Mito hi, p= 0.0005*** WT Mito hi vs NULL Mito lo), **(C)** Relative frequency of mixed colonies (n=4/sample, mean ± SD, Nested t-test, p= 0.0045** NULL Mito hi vs lo, p= 0.0029** WT Mito hi vs NULL Mito hi, p= 0.0116* WT Mito lo vs NULL Mito hi), **(D)** Relative frequency of dispersed colonies (n=4/sample, mean ± SD, Nested t-test, all comparisons not significant), **(E)** Relative colony type distribution for each sample showing proportion of compact, mixed, and dispersed colonies (n=4/sample, mean ± SD).

Amongst all of the fractions, *Kat2a* NULL Mito lo showed the least clonogenic potential which in conjunction with the observation from serial re-plating assay suggested that this cell population is difficult to sustain in *in vitro* cultures (Fig 5.3B). Further, upon looking at the clonal expansion patterns of individual samples, the ability of a single cell to expand clonally was comparable in *Kat2a* WT Mito lo and *Kat2a* NULL Mito total, again suggesting that *Kat2a* WT Mito lo may potentially phenocopy some aspects of the *Kat2a* NULL cell population (Fig 5.3B). This was also compatible with the self-renewal assay analysis discussed above. These analyses altogether highlighted that loss of *Kat2a* leads to reduction in active mitochondrial mass, which further impacts self-renewal potential and clonogenic expansion capacity of *MLL-AF9* transformed cells *in vitro*.

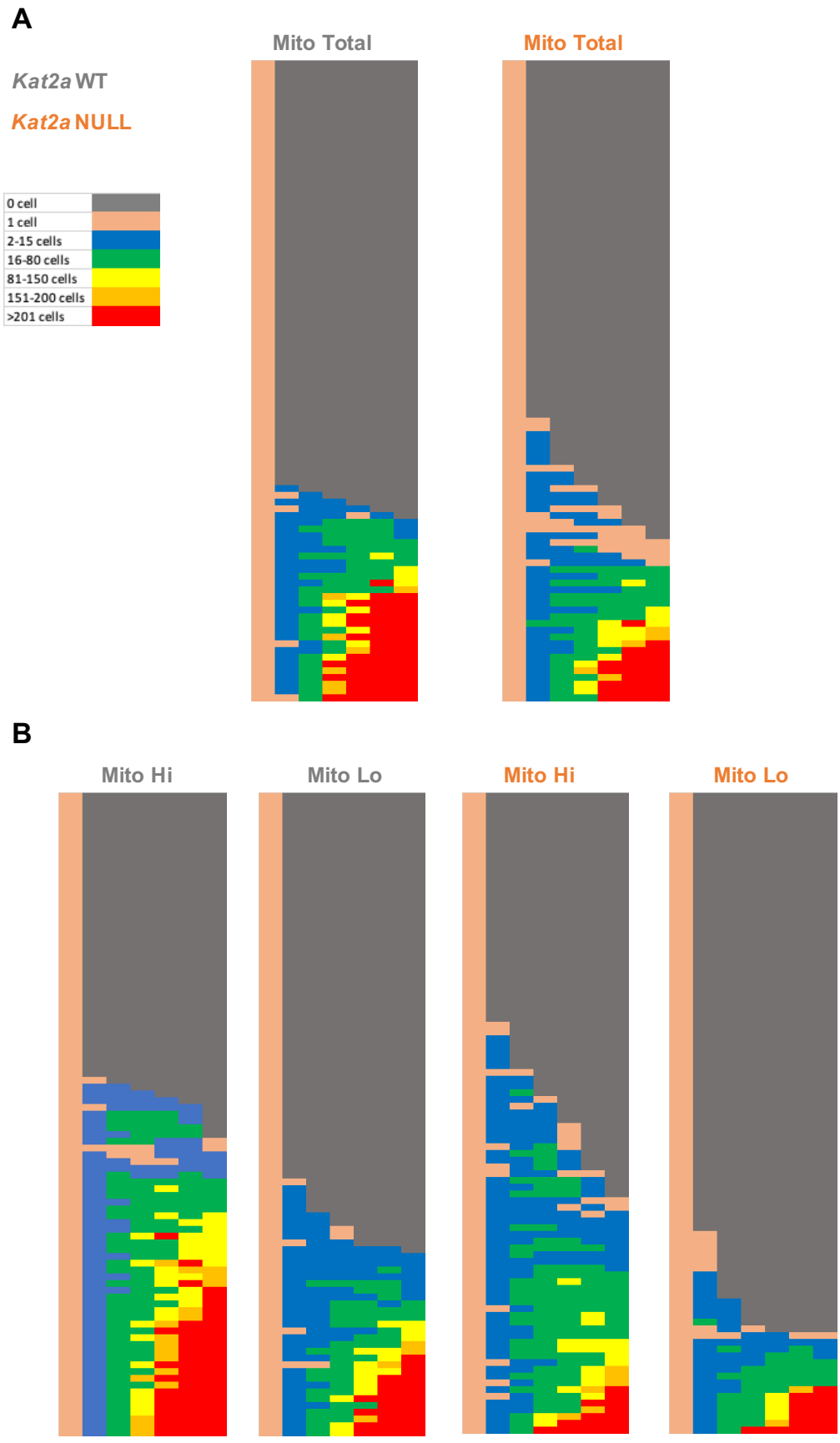

A
Kat2a WT
Kat2a NULL
0 cell
1 cell
2-15 cells
16-80 cells
81-150 cells
151-200 cells
>201 cells
Mito Total
Mito Total
B
Mito Hi
Mito Lo
Mito Hi
Mito Lo

Figure 5.3: **Single cell clonal expansion assay for *in vitro MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL BM cells sorted on the basis of mitochondrial mass.**
Heatmap representation of individual wells in a 96-well plate for maintaining single cell culture obtained from *in vitro MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL cells, **(A)** Mito total (left-*Kat2a* WT, right-*Kat2a* NULL) **(B)** Mito hi and Mito lo (left-Kat2a WT, right-*Kat2a* NULL) cells sorted on the basis of high mitochondrial mass, low mitochondrial mass, and total population (n=2/sample). The clonal expansion analysis was done on the basis of cell counting on a daily basis for 7 days.

5.3 Inhibition of mitochondrial translation in *MLL-AF9* transformed *Kat2a* WT cells phenocopies *Kat2a* NULL phenotype

Having studied the impact of *Kat2a* loss on active mitochondrial mass and consequentially on self-renewal and clonogenic expansion capacity of *MLL-AF9* transformed cells, I then studied the mechanistic association of mitochondrial translational activity on colony forming potential of these cells. Mitochondrial ribosomes differ from eukaryotic cytosolic ribosomes (O'Brien, 2003) where they use unique protein translation machinery. The process of mitochondrial translation is important for functional regulation of oxidative phosphorylation (Fukuda *et al.*, 2007). In order to study the dependency of *MLL-AF9* transformed *Kat2a* WT cells on mitochondrial translation suggestive of a consequent dependency on oxidative phosphorylation, *MLL-AF9* transformed *Kat2a* WT cells were treated with Tigecycline, an inhibitor of mitochondrial translation. Tigecycline is a third generation tetracycline antibiotic that interferes with translation by blocking the interaction of aminoacyl-tRNA with the A site of the ribosome in mitochondria (Škrtić *et al.*, 2011). Thus, tigecycline is a selective inhibitor of eukaryotic mitochondrial protein translation and respiratory complex activity.

To start the experiment, primary cell lines were generated from *MLL-AF9 in vitro* transformed *Kat2a* WT and *Kat2a* NULL cells and characterized for differences in cell cycle and apoptosis profile (Methods). After four passages, the *MLL-AF9* transformed *Kat2a* WT cells thus obtained, were treated with different concentrations of Tigecycline (2.5μM and 5μM), and a colony forming assay was performed. Colonies were scored 7 days after plating, and no

significant difference was observed in the number of colonies in Tigecycline treated *MLL-AF9* transformed *Kat2a* WT cells compared to untreated cells, suggesting that these cells might not be sensitive towards tigecycline (Fig 5.4A). However, there was a remarkable difference in the compact colonies which do not show any halo of migrating cells and represent the presence of immature myeloid cells. The reduction was more prominent in *MLL-AF9* transformed *Kat2a* WT cells treated with 5µM of tigecycline relative to the cells treated with 2.5µM tigecycline (Fig 5.4B). These observations suggested that *MLL-AF9* transformed cells may not be sensitive to tigecycline in terms of impacting the overall colony numbers, however, tigecycline treatment impacts the maintenance of self-renewal potential of these cells, highlighted by the reduced number of compact colonies. This reduction in compact colonies was accompanied by a gain in the proportion of non-compact colonies indicating that *MLL-AF9* transformed *Kat2a* WT cells upon inhibition of mitochondrial translation activity may mimic some of the characteristics of *Kat2a* NULL cells (Fig 5.4C). The non-compact colonies obtained upon tigecycline treatment had a characteristic comet shaped morphology due to which mixed and dispersed types of colonies couldn't be distinguished (Fig 5.4D).

To study any changes in the phenotype of the obtained colonies, flow cytometry analysis was conducted on both control and tigecycline treated *MLL-AF9* transformed *Kat2a* WT cells. The flow cytometry analysis revealed no significant differences in the c-Kit expression levels (marker of early progenitors) in tigecycline treated cells compared to control (Fig 5.4E). This was compatible with the global observation where no differences in number of colonies were observed upon treatment with tigecycline.

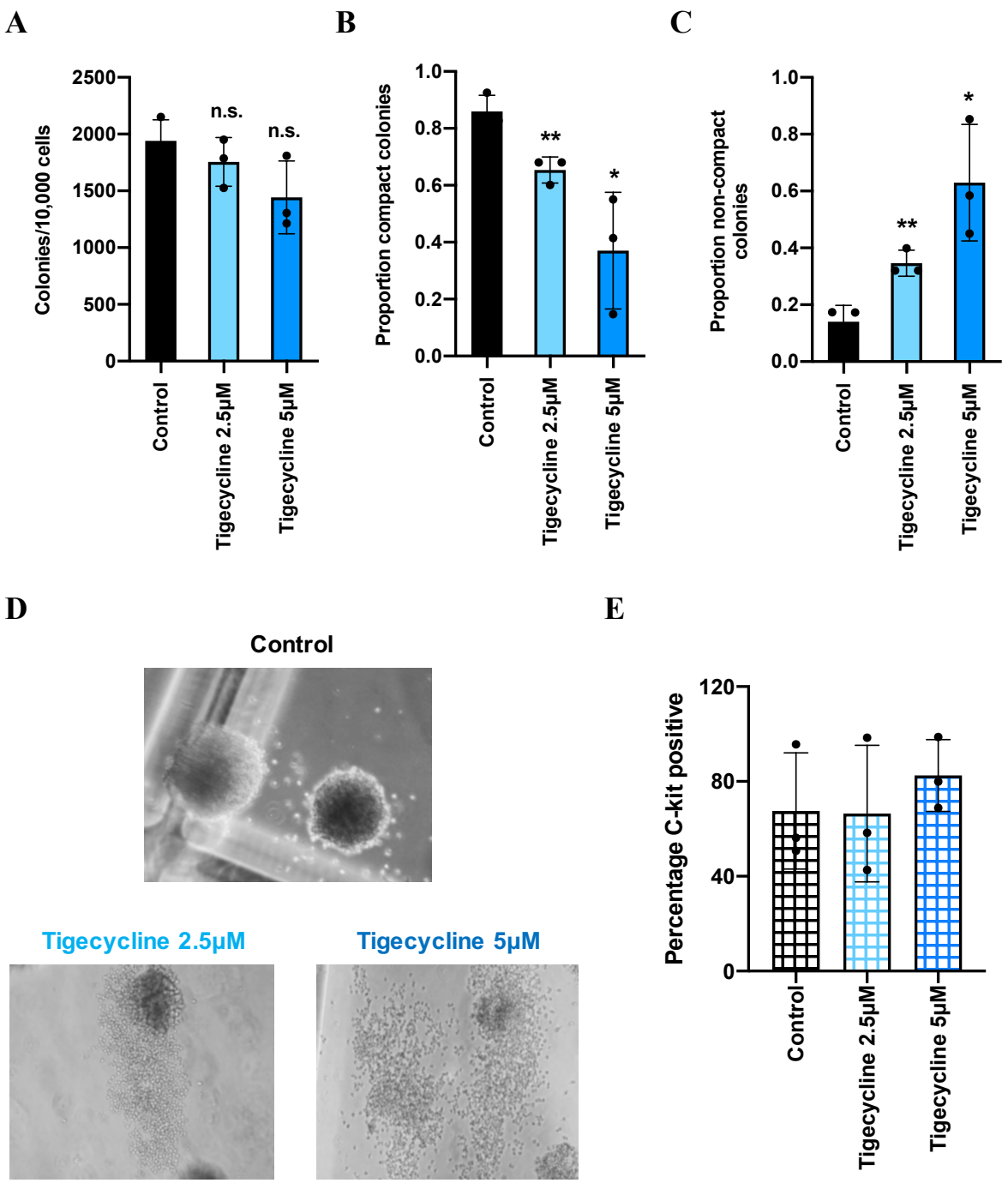

A
Colonies/10,000 cells
2500
2000
1500
1000
500
0
n.s.
n.s.
Control
Tigecycline 2.5μM
Tigecycline 5μM
B
Proportion compact colonies
1.0
0.8
0.6
0.4
0.2
0.0
**
*
Control
Tigecycline 2.5μM
Tigecycline 5μM
C
Proportion non-compact colonies
1.0
0.8
0.6
0.4
0.2
0.0
**
*
Control
Tigecycline 2.5μM
Tigecycline 5μM
D
Control
Tigecycline 2.5μM
Tigecycline 5μM
E
Percentage C-kit positive
120
80
40
0
Control
Tigecycline 2.5μM
Tigecycline 5μM

Figure 5.4: **Inhibition of mitochondrial translational activity in *in vitro MLL-AF9* transformed *Kat2a* WT BM cells.**

(A) Colony forming assay for *MLL-AF9* transformed *Kat2a* WT BM cells treated with 2.5µM and 5µM Tigecycline. Untreated cells served as control (n=3/sample, mean ± SD, Student's t-test, all comparisons non-significant), **(B)** Proportion of compact colonies (n=3/sample, mean ± SD, Student's t-test, p= 0.0091** control vs Tigecycline 2.5µM, p= 0.045* control vs Tigecycline 5µM), **(C)** Proportion of non-compact colonies (n=3/sample, mean ± SD, Student's t-test, p= 0.0091** control vs Tigecycline 2.5µM, p= 0.045* control vs Tigecycline 5µM), **(D)** Representative microscopic images of colonies at 10X resolution obtained in control, Tigecycline 2.5µM, and 5µM**.** Tigecycline treated cells showed characteristic non-compact colonies, **(E)** Flow cytometry analysis for c-Kit expression in control, Tigecycline 2.5µM, and Tigecycline 5µM treated cells (n=3/sample, mean ± SD, Student's t-test, all comparisons non-significant).

5.4 Inhibition of mitochondrial translation during *RUNX1-RUNX1T1(9a)* transformation selects for primitive cells

After studying the impact of inhibition of mitochondrial translation on reducing the self-renewal potential of *MLL-AF9* transformed *Kat2a* WT cells, I extended this analysis to *Kat2a* WT cells undergoing transformation with *RUNX1-RUNX1T1(9a)* to understand the role of mitochondrial translation in a leukaemia initiation set-up. For this, I started with *Kat2a* WT cells transduced with *RUNX1-RUNX1T1(9a)* which were maintained in a semi-solid methylcellulose medium supplemented with either 2.5µM Tigecycline or control medium in a colony forming assay. The colonies obtained at each plating were scored and analysed using flow cytometry. A reduced number of colonies were obtained at each plating upon Tigecycline treatment, suggesting that the cells undergoing transformation with *RUNX1-RUNX1T1(9a)* are sensitive to the mitochondrial translation inhibitor, Tigecycline.

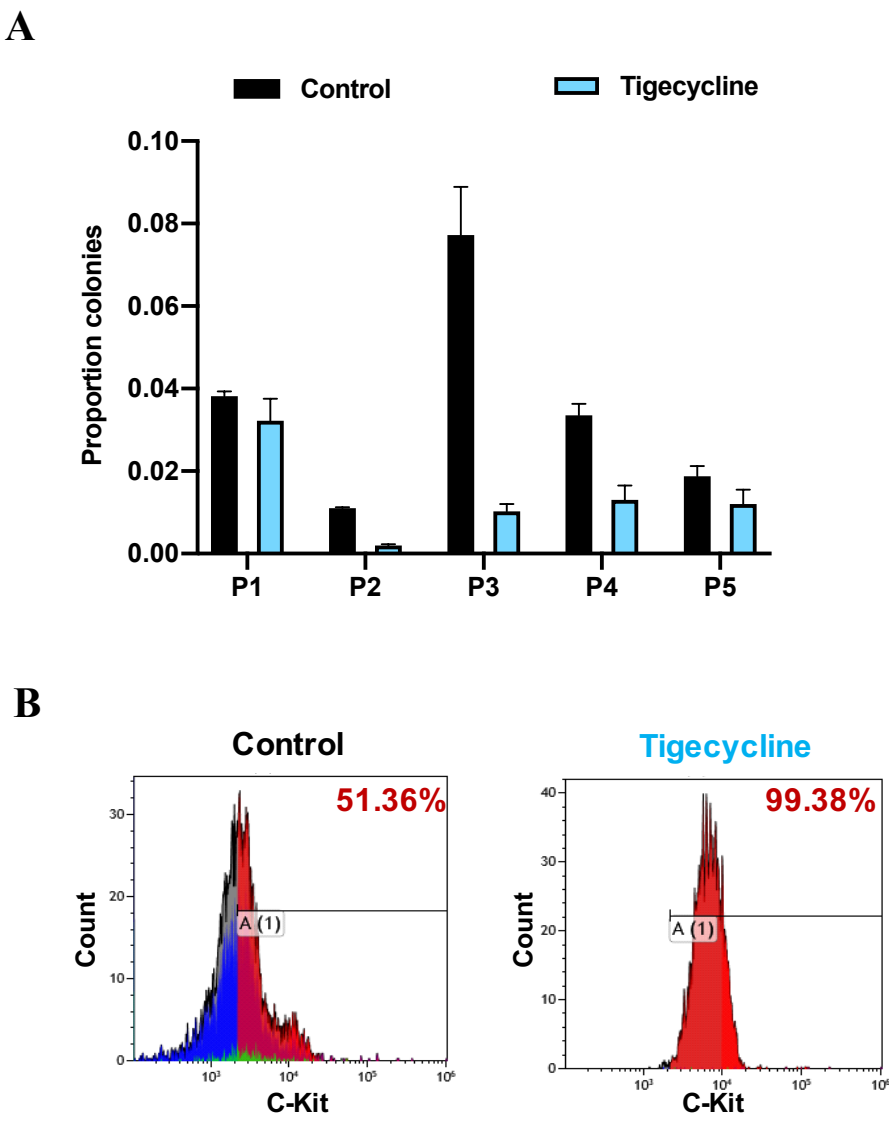


Figure 5.5: **Inhibition of mitochondrial translational activity during *RUNX1-RUNX1T1(9a)* transformation of *Kat2a* WT BM cells *in vitro*.**

(A) Proportion of colonies obtained at each plating during *RUNX1-RUNX1T1(9a)* transformation of *Kat2a* WT BM cells treated with control and Tigecycline 2.5µM (n=2/plating, mean ± SD), **(B)** Flow cytometry analysis for c-Kit expression in control and Tigecycline 2.5µM treated cells at plating 4 and 5.

Interestingly, post plating 3, which is when the percentage of *RUNX1-RUNX1T1(9a)* transformed cells is >99%, the number of colonies upon tigecycline treatment remained constant, perhaps indicative of perpetuation of a particular cell population. To characterize this cell population, I performed flow cytometry analysis, where I observed an increase in c-Kit

expression at plating 4 and plating 5, in contrast to the observation in *MLL-AF9* cells, where no changes in c-Kit expression was observed. This increase in c-Kit expression in *Kat2a* WT cells undergoing transformation with *RUNX1-RUNX1T1(9a)* suggested an accumulation of immature myeloid progenitor cells, which was contrary to the observations in *MLL-AF9* cells where an enhanced differentiation was observed upon tigecycline treatment, characterized by an increase in proportion of non-compact colonies.

Overall, both Mito hi/lo and tigecycline based approaches indicated that there are aspects of low mitochondrial activity in *Kat2a* WT cells which may capture some of the characteristics of *Kat2a* NULL cells. This was in agreement with observed transcriptional changes in *MLL-AF9* leukaemia based on scRNA-seq (Domingues *et al.*, 2020).

Having studied the role of mitochondrial biogenesis both in terms of active mitochondrial mass and mitochondrial translation in *Kat2a* WT cells transformed with *MLL-AF9* or undergoing transformation with *RUNX1-RUNX1T1(9a)*, I then studied how attenuation in ribosomal biogenesis and cytoplasmic translation machinery associated genes may impact pre-leukaemia transformation.

5.5 Loss of *Kat2a* inhibits protein synthesis during *Idh1*R132H pre-leukaemia transformation

The insights obtained from scRNA-seq analysis performed on *RUNX1-RUNX1T1(9a)* transformed pre-leukaemia cells as described in the previous chapter, indicated that loss of *Kat2a* impacts ribosomal biogenesis machinery and cytoplasmic translation processes during the process of pre-leukaemia transformation. To study the mechanistic contribution of these processes it was important to first confirm underlying differences in protein synthesis between *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells and *Kat2a* WT cells.

For this, I started with *RUNX1-RUNX1T1(9a) in vitro* transduced *Kat2a* WT and *Kat2a* NULL cells where analysis for quantification of protein synthesis was performed at each plating of the colony forming assay which maintained *RUNX1-RUNX1T1(9a) in vitro* transduced *Kat2a*

WT and *Kat2a* NULL cells in a semi-solid methylcellulose based medium enriched with cytokines. The transformed *Kat2a* WT and *Kat2a* NULL cells obtained at each plating were then utilized for OP-Puro incorporation assay to assess differences in protein synthesis between the genotypes. OP-Puro is an alkyne analogue of puromycin which is efficiently incorporated into newly translated proteins. The transformed cells obtained at each plating were briefly incubated with OP-Puro and then with AF-azide in order to detect differences based on fluorescence intensity using flow cytometry. Cells were simultaneously stained for lineage markers, Cd11b and Gr1, and the surface marker, c-Kit, (Methods) to study differences in protein synthesis in specific cell populations.

Post analysis, I observed a decrease in the OP-Puro high (Hi) population accompanied by an increase in the OP-Puro low (Lo) population of cells within $Lin^{-}Kit^{+}$ cells, in two of the biological replicates, whereas the other replicates did not indicate a reduction in protein synthesis in *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells (Fig 5.6). The contrasting results obtained with individual biological replicate made it difficult to interpret whether the indication from scRNA-seq data that *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells have reduced protein synthesis holds true in an *in vitro* set-up. It is possible that the variability observed in OP-Puro incorporation may be due to technical bias or differences in the extent of transformation on the primary cells utilised for the experiment. However, this variability could be attributed to the cellular diversity facilitated by the loss of *Kat2a* at early stages of *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation. As discussed in Chapter-6 in more detail, loss of *Kat2a* specifically at the 2 months time point post transplantation showed cellular differentiation towards monocytic and B-cell lymphocyte lineage. This cellular variability could lead to differential OP-Puro incorporation and lead to ambiguous results. Although *RUNX1-RUNX1T1(9a)* transformed cells didn't show a significant change in overall protein synthesis upon loss of *Kat2a*, there was an indication in two of the biological replicates towards reduced protein synthesis in *Kat2a* NULL cells. To investigate this further, I looked at OP-Puro incorporation in *Idh1*R132H model.

For this, *Idh1*R132H cells with either *Kat2a* HET or *Kat2a* NULL genotype processed 20 weeks post *pIpC* treatment were thawed and maintained in a serial re-plating assay (Methods). Similar to the *RUNX1-RUNX1T1(9a)* methodology, the transformed *Idh1*R132H *Kat2a* HET

and *Idh1*R132H *Kat2a* NULL cells obtained at each plating were utilized for OP-Puro incorporation assay to assess differences in protein synthesis between the genotypes. The cells obtained were briefly incubated with OP-Puro and AF-azide as mentioned previously and analysed for OP-Puro incorporation. A significant reduction in OP-Puro high (Hi) population was observed within the $Lin^{-}c\text{-}Kit^{+}$ compartment of cells in *Kat2a* NULL cells transformed with *Idh1*R132H (Fig 5.7A and 5.7B). This was also accompanied by a gain in OP-Puro low (Lo) population (Fig 5.7C), suggesting that loss of *Kat2a* reduces translation rate in *Idh1*R132H pre-leukaemia cells. In contrast to *RUNX1-RUNX1T1(9a)* model, a combined analysis of all of the biological replicates in *Idh1*R132H model, reached statistical significance (Fig 5.7B and 5.7C), suggesting that perturbation of ribosomal biosynthetic programmes may play a mechanistic role during pre-leukaemia transformation. These findings were compatible with reduced protein synthesis observed upon *Kat2a* loss in *MLL-AF9* leukaemia (Domingues *et al.*, 2020).

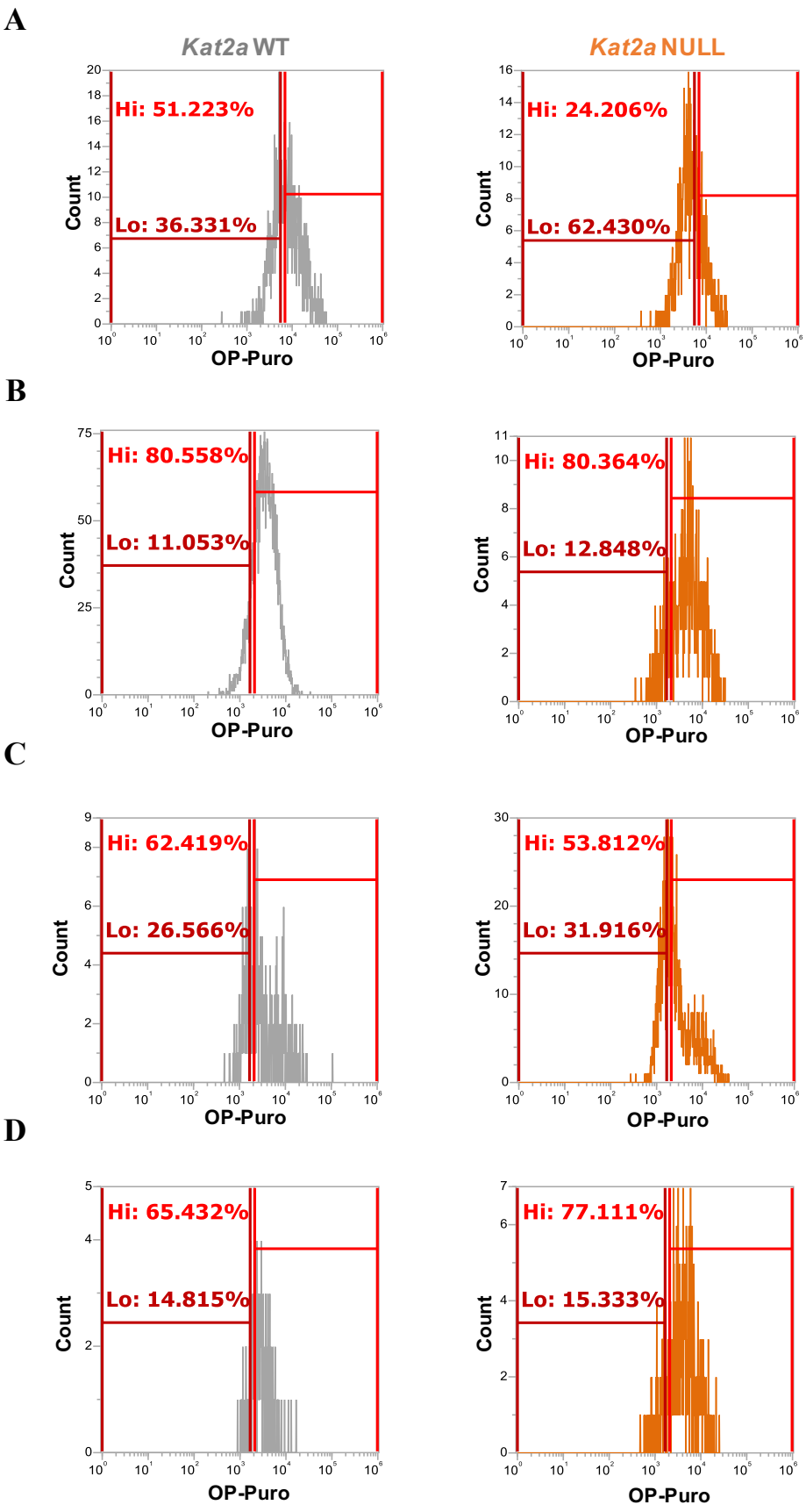

A
Kat2a WT
Hi: 51.223%
Lo: 36.331%
Count
OP-Puro
Kat2a NULL
Hi: 24.206%
Lo: 62.430%
Count
OP-Puro
B
Hi: 80.558%
Lo: 11.053%
Count
OP-Puro
Hi: 80.364%
Lo: 12.848%
Count
OP-Puro
C
Hi: 62.419%
Lo: 26.566%
Count
OP-Puro
Hi: 53.812%
Lo: 31.916%
Count
OP-Puro
D
Hi: 65.432%
Lo: 14.815%
Count
OP-Puro
Hi: 77.111%
Lo: 15.333%
Count
OP-Puro

Figure 5.6: **OP-Puro analysis for *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT and *Kat2a* NULL BM cells.**

Flow cytometry histogram plots for *Kat2a* WT (left) and *Kat2a* NULL (right) BM cells highlighting OP-Puro high (Hi) and OP-Puro low (Lo) cell populations, **(A)** Biological replicate 1, **(B)** Biological replicate 2, **(C)** Biological replicate 3 **(D)** Biological replicate 4.

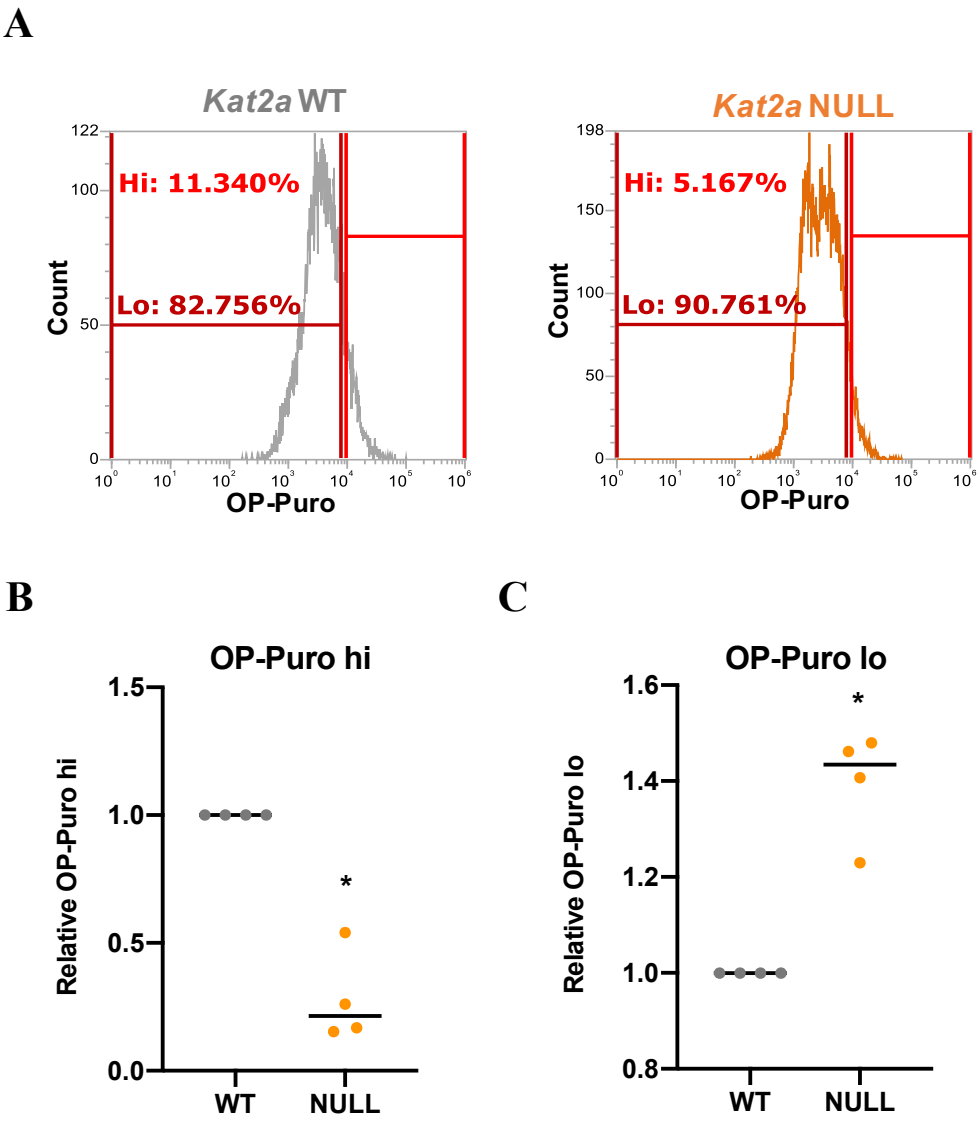


Figure 5.7: OP-Puro analysis for *Idh1*R132H transformed *Kat2a* WT and *Kat2a* NULL BM cells. **(A)** Representative flow cytometry histogram plots for *Kat2a* WT (left) and *Kat2a* NULL (right) BM cells highlighting OP-Puro high (Hi) and OP-Puro low (Lo) cell populations, **(B)** Plot representing relative OP-Puro high population in *Idh1*R132H transformed *Kat2a* WT and *Kat2a* NULL BM cells (n=4/sample, Nested t-test, p= 0.0267*), **(C)** Plot representing relative OP-Puro low population in *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT and *Kat2a* NULL BM cells (n=4/sample, Nested t-test, p= 0.0353*).

The reduced incorporation of OP-Puro in different models of AML, including *RUNX1-RUNX1T1(9a)*, *Idh1*R132H and *MLL-AF9* models of leukaemia strengthen previous findings from transcriptomics data that loss of *Kat2a* is associated with reduced protein synthesis and the transcriptional programmes associated with ribosomal biogenesis and protein synthesis may play a mechanistic role in *Kat2a* mediated pre-leukaemia transformation.

5.6 Inhibition of protein synthesis in *RUNX1-RUNX1T1(9a)* and *Idh1*R132H model aids in pre-leukaemia transformation

Having confirmed that loss of *Kat2a* is associated with reduced protein synthesis, I wanted to understand the mechanistic contribution of attenuation of ribosomal biosynthetic programmes during pre-leukaemia transformation. To study this, I treated *RUNX1-RUNX1T1(9a) in vitro* transduced *Kat2a* WT cells with S6K1 inhibitor (control-DMSO) during colony forming assay (Methods). The colonies obtained were re-plated in the presence of S6K1 inhibitor or control DMSO until three platings. S6K1 inhibitor acts on S6K1 isoform of p70 ribosomal S6 kinase and thus inhibits protein synthesis. To mediate its effects on metabolic pathways, S6K1 phosphorylates a number of downstream substrates including the small ribosomal subunit protein S6 (rpS6) (Salmond *et al.*, 2015).

To confirm the activity of S6K1 inhibitor, OP-Puro assay was conducted to assess differences in protein synthesis upon S6K1 inhibition. The *Kat2a* WT cells undergoing transformation with *RUNX1-RUNX1T1(9a)* were treated with 10mM of S6K1 inhibitor with DMSO serving as a control. OP-Puro incorporation analysis was done where the cells treated with S6K1 inhibitor showed a reduction in OP-Puro high (Hi) population with a significant gain in OP-

Puro low (Lo) population, thus validating the reduction in rate of translation upon S6K1 inhibition. (Fig 5.7A). This trend was consistent across all biological replicates (Fig 5.7B), confirming the functional output of S6K1 inhibitor on *Kat2a* WT cells.

After confirming the functionality of S6K1 inhibitor, the colonies representative of self-renewal capacity were scored after each plating. *Kat2a* WT cells transformed with *RUNX1-RUNX1T1(9a) in vitro* were found to have enhanced colony numbers representing an increased self-renewal capacity upon S6K1 inhibition compared to control at plate 2 (Fig 5.8A). Contrastingly, there was no difference in the number of colonies at plate 3, indicating that reduced protein synthesis may aid in early *RUNX1-RUNX1T1(9a)* transformation (Fig 5.8B).

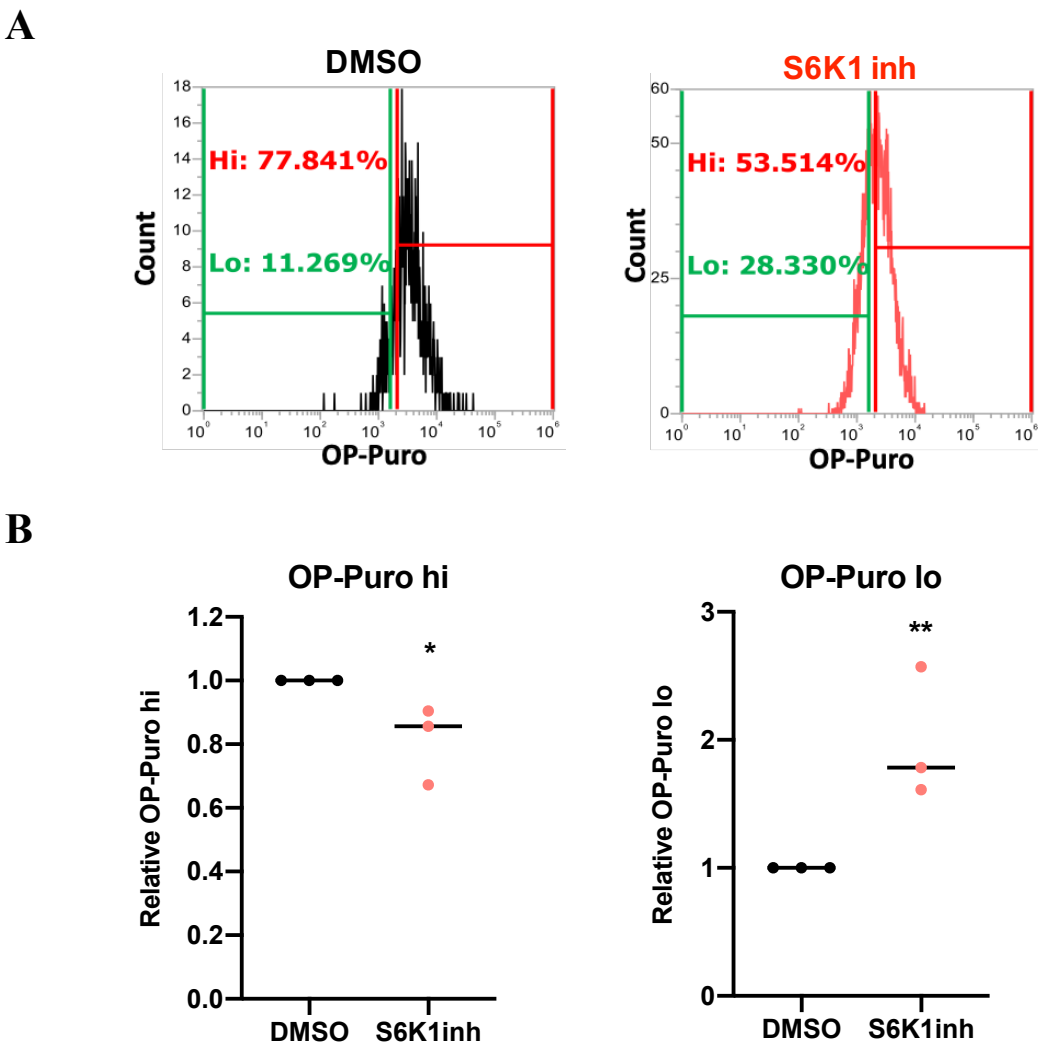


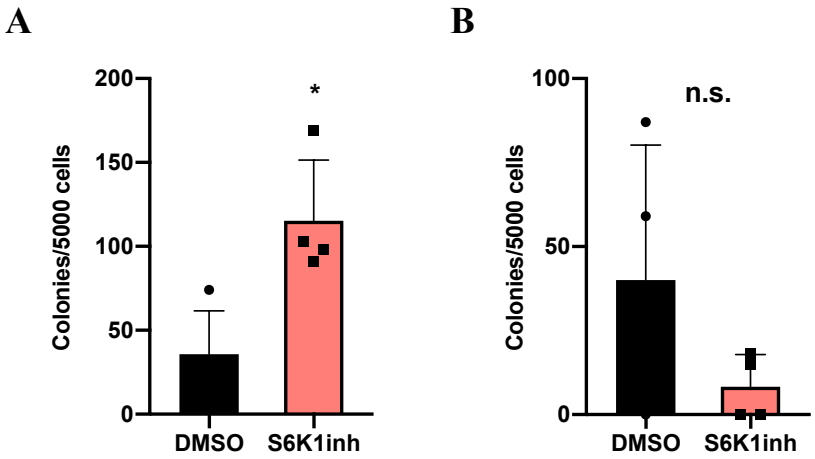
Figure 5.8: **OP-Puro analysis for *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT cells treated with S6K1 inhibitor.**

(A) Representative flow cytometry plot for OP-Puro analysis of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT BM cells treated with control-DMSO (left) and treated with S6K1 inhibitor (right). OP-Puro high (Hi) (pink) and OP-Puro low (lo) (green) populations are highlighted, **(B)** Relative OP-Puro hi population (left) and OP-Puro lo population (right) upon S6K1 inhibitor treatment (n=3/sample for each comparison, Nested t-test, p= 0.0242* for OP-Puro hi, p= 0.0054** for OP-Puro lo). Horizontal bars represent mean value and each dot represents an individual biological replicate.

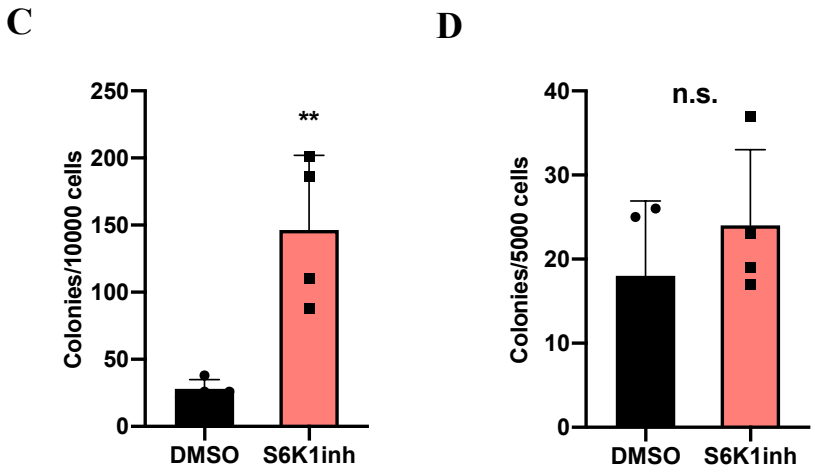
To test whether this observation is model specific, similar analysis was done using *Idh1*R132H *Kat2a* WT cells processed 4 weeks post *pIpC*. There was a significant association of enhanced colony forming efficiency at plate 2 upon S6K1 inhibition (Fig 5.8C), which disappeared at plate 3 (Fig 5.8D), overall strengthening the findings in *RUNX1-RUNX1T1(9a)* that inhibition of protein synthesis aids in pre-leukaemia transformation which further contributes to accelerated leukaemic progression.

Having observed the mechanistic contribution of perturbation of protein synthesis in pre-leukaemia models, I further investigated its role in the *MLL-AF9* leukaemia model. *Kat2a* WT cells transformed with *MLL-AF9 in vitro* were treated with S6K1 inhibitor and I observed that there was a significant reduction in colony forming efficiency overall, indicating a reduction in self-renewing capacity of these cells compatible with the phenotype observed in *MLL-AF9* transformed *Kat2a* NULL cells (Domingues *et al.*, 2020) (Fig 5.8E). The S6K1 inhibition also led to a reduction in compact colonies representative of immature myeloid cells, accompanied by an increase in proportion of mixed and dispersed colonies, representative of differentiated macrophages, again compatible with the *MLL-AF9* transformed *Kat2a* NULL phenotype (Fig 5.8F). These findings overall suggested a mechanistic role for ribosomal biogenesis and protein synthesis machinery associated genes in pre-leukaemic transformation. However, the perturbation is context dependent, where reduced expression in pre-leukaemia models accelerates pre-leukaemia transformation, whereas in a maintenance model, it leads to enhanced cellular differentiation.

RUNX1-RUNX1T1(9a)



***Idh1*R132H**



MLL-AF9

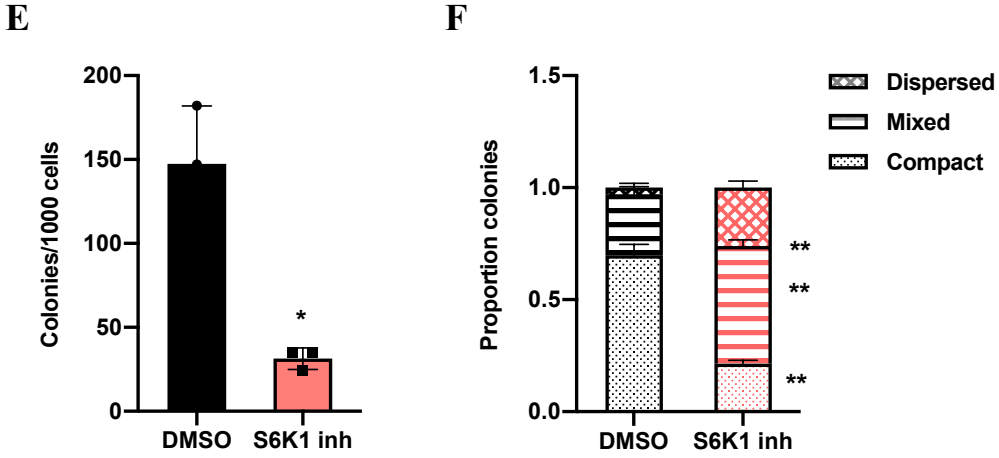


Figure 5.9: **Colony forming assay upon S6K1 inhibition.**
Number of colonies obtained during serial re-plating of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT BM cells treated with S6K1 inhibitor where DMSO served as control at **(A)** Plate 2 (n=4/sample, mean ± SD, Nested t-test, p= 0.0117*), **(B)** Plate 3 (n=4/sample, mean ± SD, Nested t-test, p= 0.1753 n.s.), Number of colonies obtained during serial re-plating of *Idh1*R132H transformed *Kat2a* WT BM cells treated with S6K1 inhibitor where DMSO served as control at **(C)** Plate 2 (n=4/sample, mean ± SD, Nested t-test, p= 0.0056**), **(D)** Plate 3 (n=4/sample, mean ± SD, Nested t-test, p= 0.3803 n.s.), **(E)** Number of colonies obtained when *MLL-AF9* transformed *Kat2a* WT BM cells were treated with S6K1 inhibitor. DMSO treatment served as control (n=3/sample, mean ± SD, Student's t-test, p= 0.0251*), **(F)** Proportion of colonies obtained when *MLL-AF9* transformed *Kat2a* WT BM cells were treated with S6K1 inhibitor. DMSO treatment served as control (n=3/sample, mean ± SD, Student's t-test, p= 0.0016** for compact colonies, p= 0.0034** for mixed colonies and p= 0.0048** for dispersed colonies).

In this chapter, I discussed the mechanistic association of transcriptional programmes, namely mitochondrial bioenergetics, mitochondrial translation, and ribosomal biogenesis/cytoplasmic translation which were found to be impacted upon loss of *Kat2a* during *RUNX-RUNX1T1(9a)* pre-leukaemia transformation. Although mitochondrial bioenergetics mechanism may highlight the initial metabolic reconfiguration as a consequence of *Kat2a* loss, at the early stages of pre-leukaemia transformation, I hypothesized that ribosomal biogenesis and cytoplasmic translational machinery may reflect the mechanistic association with *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation.

I started by looking at the differences in mitochondrial activity in terms of active mitochondrial mass and mitochondrial potential in *RUNX-RUNX1T1(9a)* transformed pre-leukaemia *Kat2a* WT and *Kat2a* NULL cells. Since *MLL-AF9* transformed cells could aid in leukaemia maintenance by immortalization of a leukaemogenic cell population which is dependent on constitutive *MLL-AF9* expression, I utilized this as a contrasting model to *RUNX1-RUNX1T1(9a)* pre-leukaemia model. The reduction in active mitochondrial mass upon loss of *Kat2a* was synergistic in both models, confirming the scRNA-seq insights that *Kat2a* loss is associated with reduction in mitochondrial activity. Surprisingly, no differences were observed in terms of mitochondrial potential indicating that the reduction in active mitochondrial mass

was subtle and unable to induce any consequent difference in mitochondrial potential. In the future, this analysis may be extended by looking at the levels of oxidative phosphorylation in terms of the presence of reactive oxygen species. Although the findings were compatible in both *RUNX-RUNX1T1(9a)* pre-leukaemia as well as *MLL-AF9* leukaemia model, the outcome in both models upon reduction in mitochondrial activity is different, suggesting that loss of *Kat2a* is merely facilitating rather than determining the fate acquisition which consequently leads to its model-specific phenotype.

In order to study the impact of mitochondrial activity on the *MLL-AF9* leukaemia transformation as well as on self-renewal, the *Kat2a* WT and *Kat2a* NULL cells transformed with *MLL-AF9* were segregated on the basis of active mitochondrial mass and the different cell populations obtained post-sorting were maintained in a colony forming assay for a week. The colonies obtained suggested a reduction in colony forming potential in Mito lo fraction of cells irrespective of the genotype, indicating that reduced mitochondrial mass may impact the self-renewal potential of *MLL-AF9* cells, which could be further extrapolated in other models of leukaemia. The reduction in colonies obtained in *Kat2a* WT Mito lo accompanied by reduced compact colonies suggested that *Kat2a* WT Mito lo may phenocopy *Kat2a* NULL cells as a whole. These experiments were further extended in the bone marrow cells obtained from secondary leukaemia, however, an abrupt increase in colony forming efficiency was observed in *Kat2a* NULL Mito hi, perhaps due to variability in primary samples, due to which the experiment could not be interpreted.

The experiments on self-renewal were accompanied by the single cell clonal assay where individual cells were index sorted from Mito hi and Mito lo cell populations obtained from individual genotypes and were assessed for their clonal expansion capacity. The analysis was in line with the results obtained from colony forming assay, where cells obtained from Mito lo subfraction had reduced clonal expansion capacity compared to their respective Mito hi fractions. Once again, the pattern of clonogenicity observed in *Kat2a* WT Mito lo phenocopied that of *Kat2a* NULL total cells. It would have been interesting to study the phenotype of individual clones obtained from these cell fractions using flow cytometry in order to correlate their phenotype with *MLL-AF9* leukaemic animals. However, due to limited *in vitro* expansion

capacity of *RUNX1-RUNX1T1(9a)* cells, these assays could not be replicated in a leukaemia initiation set up.

Another aspect of mitochondrial bioenergetics was to study the impact of mitochondrial translation on leukaemia transformation and self-renewal capacity. To understand the role of mitochondrial translation during *MLL-AF9* leukaemia, the *Kat2a* WT cells transformed with *MLL-AF9* were treated with Tigecycline, an inhibitor of mitochondrial translation. The inhibition of mitochondrial translation using tigecycline suggested no difference in global colony forming capacity, in contrast to the findings made from Mito/Mito lo assay. These contrasting observations could be due to the translation aspect of mitochondrial activity. With literature suggesting that AML cells with lower mitochondrial mass are generally resistant to tigecycline (Škrtić *et al.*, 2011), it might be worth performing the tigecycline treatment on the cell populations segregated on the basis of active mitochondrial mass. This may also be valid in case of *RUNX1-RUNX1T1(9a)* where a selection of tigecycline resistant colonies was observed upon treatment with tigecycline. These cells however, represented an early progenitor population, unlike *MLL-AF9* transformed cells, suggesting perpetuation of early progenitors in a reduced mitochondrial translation environment, in line with findings from scRNA-seq data. However, this analysis could be repeated with a greater number of biological replicates in order to draw stronger conclusions. It would have been interesting to study the role of mitochondrial translation in *Idh1*R132H pre-leukaemia transformation process given the association of *Idh1* mutation with accumulation of 2-hydroxyglutarate (2-HG) which consequently leads to metabolic reprogramming in AML. Inhibition of mitochondrial translation in this model would also impact the process of oxidative phosphorylation during tricarboxylic acid cycle (TCA) which would affect the metabolic landscape and may further impact the process of pre-leukaemia transformation.

After studying the role of mitochondrial activity and mitochondrial translation in pre-leukaemia transformation (*RUNX1-RUNX1T1(9a)*) and leukaemia progression (*MLL-AF9*), I studied the impact of attenuation in translational machinery on these processes. The OP-Puro incorporation assay performed to assess any differences in protein synthesis rate upon *Kat2a* loss suggested a reduction in case of *Idh1*R132H model, contrary to the observations in *RUNX1-RUNX1T1(9a)* where no significant differences were observed globally, except for a

slight trend towards reduction in OP-Puro high population in two of the replicates. I attributed this variability in OP-Puro incorporation to the enhanced cellular diversity generated upon *Kat2a* loss (Chapter-6). It is worth considering that this variability in OP-Puro incorporation wasn't observed in case of *Idh1*R132H model, as the cells utilized for OP-Puro incorporation assay were not at very early stage of pre-leukaemia transformation, at which time point the cellular diversity was observed in case of *RUNX1-RUNX1T1(9a)* model. Future analysis may involve studying the incorporation of OP-Puro in *RUNX1-RUNX1T1(9a) Kat2a* WT and *Kat2a* NULL animals, which would enable an in-depth analysis of alterations in protein synthesis levels during pre-leukaemia progression.

Another approach to study the role of ribosomal biogenesis during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation was to make use of shRNAs against Rpl38 gene, due to its potential contribution to Hox*a* programmes (Kondrashov *et al.*, 2011b), which could reflect a visible difference between *MLL-AF9* leukaemia and *RUNX1-RUNX1T1(9a)* pre-leukaemia set-up. For this, the *Kat2a* WT cells transduced with *RUNX1-RUNX1T1(9a)* were infected with two different shRNAs targeting Rpl38 gene. Post infection, 18-20% of the cells were found to be positive for Rpl38 shRNAs as well as control (mCherry reporter). Unfortunately, these transduced cells could not be maintained successfully in a colony forming assay set-up, perhaps due to the incapability of Rpl38sh cells in promoting expansion of early progenitors (Kondrashov *et al.*, 2011b).

After seeing the differences in protein synthesis rate in *Kat2a* NULL cells transformed with *Idh1*R132H, compatible with previous lab findings in *MLL-AF9* model (Domingues *et al.*, 2020), I sought to associate these findings mechanistically by performing inhibition of protein synthesis using S6K1 inhibitor which acts on S6K1 isoform of p70 ribosomal S6 kinase and thus inhibits protein synthesis. Upon performing S6K1 inhibition, an enhanced colony forming potential was observed at plating 2, in line with the insights from *RUNX1-RUNX1T1(9a)* scRNA-seq analysis, suggesting that impairment of ribosomal biogenesis/cytoplasmic translation may contribute towards the acceleration in pre-leukaemia transformation. This increase in colony forming potential was lost markedly upon plating 3 indicating that inhibition of cytoplasmic translation is beneficial for *RUNX1-RUNX1T1(9a)* pre-leukaemia progression during early stages of transformation. Similar findings were observed in case of *Idh1*R132H,

where an enhanced colony forming potential was observed at plating 2 with no difference at plating 3. However, in case of *MLL-AF9* model, reduced colony forming potential was observed accompanied by a decrease in compact colonies and an increase in mixed and dispersed colonies, in line with the scRNA-seq analysis performed on *MLL-AF9* leukaemia *Kat2a* WT and *Kat2a* NULL cells (Domingues *et al.*, 2020). These findings altogether suggest that inhibition of cytoplasmic translation in *Kat2a* WT cells phenocopies some aspects of *Kat2a* NULL cells.

Although the findings from S6K1 inhibition experiment were compatible with scRNA-seq insights obtained from individual models at leukaemia and pre-leukaemia stages, however, it represents a dichotomy since inhibition of cytoplasmic translation promotes pre-leukaemia transformation in *RUNX1-RUNX1T1(9a)* model, whereas it leads to inhibition of leukaemia stem-like cells in case of *MLL-AF9* model. These findings reflect that reduction in protein synthesis is specifically associated with loss of *Kat2a*, which facilitates rather than determines the cellular state during the process of leukaemia progression. The next chapter in this thesis reflects on how loss of *Kat2a* promotes cellular diversity by increasing cell-to-cell transcriptional variability in *RUNX1-RUNX1T1(9a)* pre-leukaemia cells.

6 Analysis of the role of transcriptional variability upon *Kat2a* loss in *RUNX1-RUNX1T1(9a)* pre-leukaemia

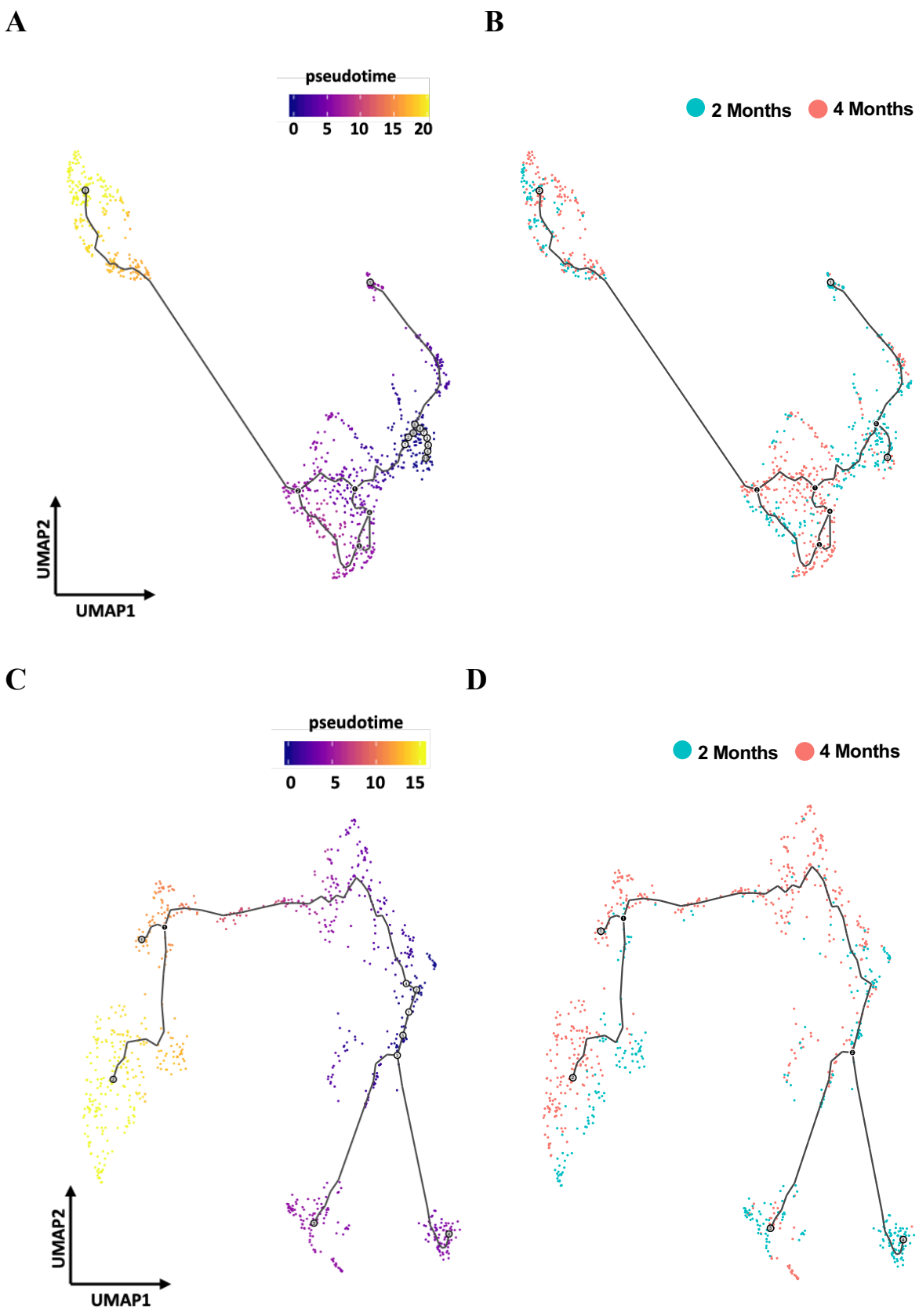
The previous chapters discussed the functional impact of *Kat2a* loss in *RUNX1-RUNX1T1(9a)* as well as *Idh1*R132H pre-leukaemia models, where loss of *Kat2a* accelerates leukaemia progression along with perpetuation of pre-leukaemia clones characterized by enhanced self-renewal capacity. These pre-leukaemia cells benefit from a reduced translational activity as well as reduced mitochondrial biosynthetic programmes during different stages of pre-leukaemia progression and aid in leukaemia development. The process of pre-leukaemia development, similar to other developmental processes, require continuous changes in the trajectory of a transformed cell. It would be ideal to monitor these changes in a given cell over the period of pre-leukaemia progression. Unfortunately, as the cells subjected to single cell RNA sequencing (scRNA-seq) are lysed (destroyed) when the RNA is extracted, we are limited to studying cell fate trajectory by sampling at multiple time points and obtaining snapshots of the gene expression profile of individual cells. Since cells not having *Kat2a* should proceed faster along the pre-leukaemia development trajectory than other cells, each snapshot may contain cells at varying points along the developmental progression. This information can therefore be analysed using different statistical methods to order cells along one or more trajectories which represent the underlying developmental trajectories. Such an ordering is often termed as "pseudotime". As mentioned in Chapter-1, an underlying cause of cell fate transitions could be the presence of inherent cell-to-cell transcriptional variability. Given the central role of *Kat2a* in limiting cell-to-cell transcriptional variability (Moris *et al.*, 2018b) (Domingues *et al.*, 2020), I further interrogated a potential link between loss of *Kat2a*, its consequent increase in transcription variability, and pre-leukaemia progression, by utilising the single-cell transcriptomics data discussed in Chapter-4.

6.1 Single- cell pseudotime trajectory analysis highlights that *Kat2a* NULL cells follow a dispersed trajectory

To study the differences between transcriptional dynamics of *Kat2a* WT and *Kat2a* NULL cells, I used Monocle 3 (v3.0) (Trapnell *et al.*, 2014) to build trajectories based on time series

progression data. Single-cell trajectory building was important to capture asynchrony between the population of cells of a genotype. Since Monocle 3 works by grouping cells with similar gene expression profiles at early stages, I pre-processed the data separately keeping the filtering parameters similar to the ones used in Seurat v2.4 analysis (Chapter-4). To start, I created a cell data set (CDS) object from raw data containing the expression matrix, cell metadata, and gene annotations. After this, I normalized the data by log transformation and size factors to address depth differences. Post normalization, I did PCA in order to study the variance contributed by each Principal Component (PC) (Methods). I visualized cells in a lower dimensional space using Uniform Manifold Approximation and Projection (UMAP) which is implemented in Monocle 3 (Methods). Single-cell trajectory building was done by ordering each cell captured in scRNA-seq according to its progress along a learned pseudotime trajectory. Pseudotime represents the distance between a cell from the start of the trajectory, measured along the shortest path. The total length of the trajectory is defined in terms of the total amount of transcriptional change that a cell undergoes as it moves from the starting state to the end state.

To start, I built a single-cell trajectory for *Kat2a* WT cells processed at both 2 months and 4 months post transplantation. For this, I first visualized the cells using UMAP dimensionality reduction method. Then, I performed pseudotime temporal ordering (Fig 6.1A) with the root node of the trajectory set to the space occupied by *Kat2a* WT cells processed at 2 months (Fig 6.1B). The trajectory started from these cells (1-10, white) (Fig 6.1A) and was followed by a bifurcation. Going forward, one of the branches is referred to as cellular state 1 (1, grey). The other branch which undergoes transcriptional modifications (1-4, black) and subsequently leads to a state which is referred to as cellular state 2 (2, grey) (Fig 6.1A). Overall, the *Kat2a* WT cells followed a linear trajectory starting from the root (purple) and ending up at the final (yellow) cellular state (Fig 6.1A). The branch defined by cellular state 1 was enriched in *Kat2a* WT cells at 2 months, perhaps indicating an early cell fate during *RUNX1-RUNX1T1(9a)* transformation, whereas cellular state 2 was composed of cells from both the 2 months and 4 months timepoints (Fig 6.1B).


A
B
pseudotime
0 5 10 15 20
2 Months
4 Months
UMAP2
UMAP1
C
D
pseudotime
0 5 10 15
2 Months
4 Months
UMAP2
UMAP1

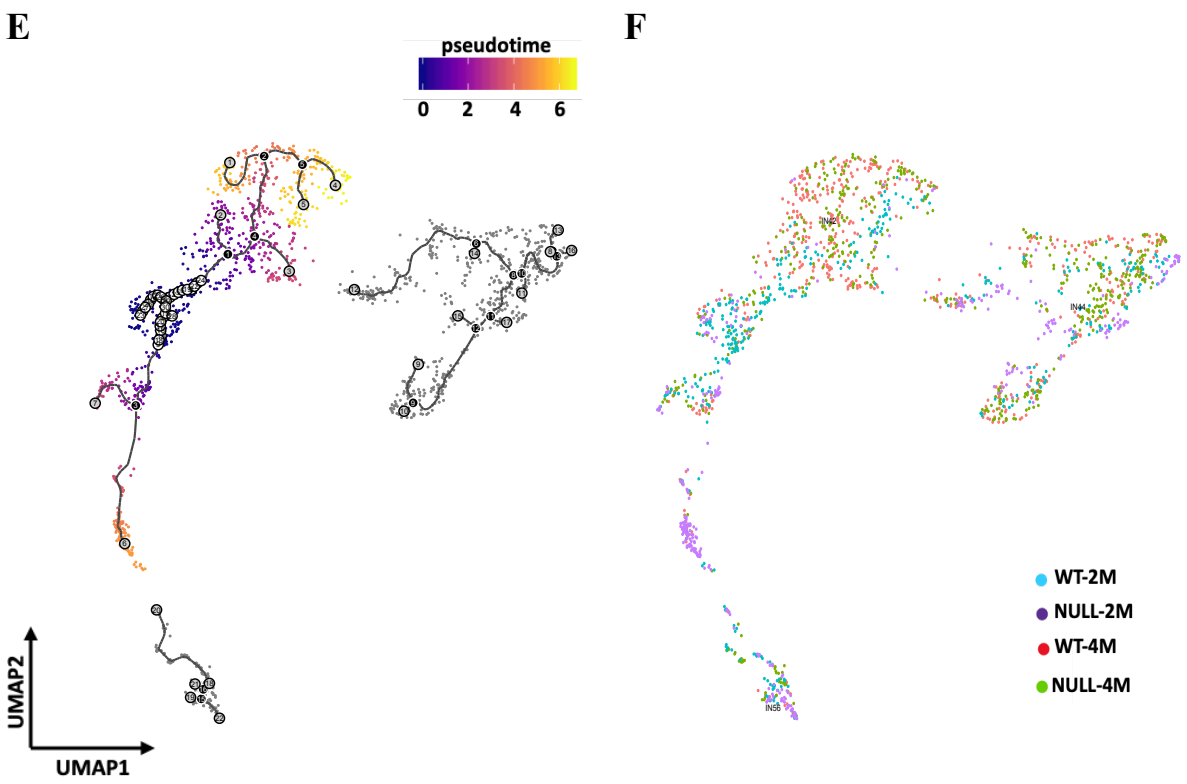


Figure 6.1: Pseudotime trajectory analysis for single cells using Monocle 3.0.

(A) Pseudotime trajectory represented on a UMAP plot for *Kat2a* WT cells, including those collected at both 2 months and 4 months post transplantation. The encircled cells (white) represent the start of the trajectory highlighted in purple colour, whereas yellow-coloured cells represent the end of the trajectory. Cells encircled in black represent a different cellular state, whereas cells encircled in grey represent branching, **(B)** UMAP representation of *Kat2a* WT cells at both time points of 2 months (blue) and 4 months (pink) post transplantation, **(C)** Pseudotime trajectory representation on a UMAP plot for all *Kat2a* NULL cells including those collected at both 2 months and 4 months timepoints, **(D)** UMAP representation of *Kat2a* NULL cells at individual time points of 2 months (blue) and 4 months (pink) post transplantation, **(E)** Global pseudotime trajectory represented on a UMAP plot for all *Kat2a* WT and *Kat2a* NULL cells including both 2 months and 4 months timepoints, **(F)** Global UMAP representation of all *Kat2a* WT and *Kat2a* NULL cells at both time points post transplantation, *Kat2a* WT 2 months (blue), *Kat2a* WT 4 months (red), *Kat2a* NULL 2 months (purple) and *Kat2a* NULL 4 months (green).

Similarly, I built a trajectory for *Kat2a* NULL cells at both 2 months and 4 months post transplantation. Pseudotime temporal ordering was done (Fig 6.1C) post UMAP visualization, with the root node set to *Kat2a* NULL cells processed at 2 months (1-5, white) (Fig 6.1C).

Unlike *Kat2a* WT cells, *Kat2a* NULL cells had a branched trajectory with four different cellular states (1-4, grey). One of these branches was composed of *Kat2a* NULL cells at 2 months exclusively, indicating a cellular state consequential to loss of *Kat2a* during early stages of *RUNX1-RUNX1T1(9a)* transformation, whereas the rest of the cellular states were composed of *Kat2a* NULL cells from both the 2 months and 4 months timepoints (Fig 6.1D).

After studying differences in individual trajectories in both *Kat2a* WT and *Kat2a* NULL cells, a global trajectory was built in a similar manner, including all *Kat2a* WT and *Kat2a* NULL cells at both time points (Fig 6.1E). Pseudotime temporal ordering was performed with the root node set to the space occupied by both *Kat2a* WT and *Kat2a* NULL cells processed at 2 months (1-30, white) (Fig 6.1F). Again, a bifurcation was observed from the root cells (Fig 6.1E), where one of these branches was enriched for *Kat2a* NULL cells at 2 months and *Kat2a* WT cells at 4 months, and the other branch was composed of *Kat2a* NULL cells at 2 months exclusively (Fig 6.1F). This was compatible with the observations from the genotype specific trajectories. However, a discontinuity was observed within the trajectory, which may be the consequence of an abrupt leukaemia transformation event. This discontinuity in global trajectory was in-line with the abrupt transition of cellular state 2 in the *Kat2a* WT specific trajectory. Cell-to-cell correspondence analysis between the genotype-specific and global trajectories indicated that the population of cells characterized by cellular state 2 in the individual genotype trajectories corresponded to the subset of cells enriched in *Kat2a* NULL cells at 2 months and *Kat2a* WT cells at 4 months in the global trajectory. Overall, trajectory analysis indicated presence of a branched trajectory upon loss of *Kat2a,* suggestive of transcriptional reprogramming at various steps during pre-leukaemia progression contributing towards a changing cellular landscape.

6.2 Single-cell trajectory coincides with haematopoietic hierarchy

Having identified differences in progression of *Kat2a* WT and *Kat2a* NULL cells, I wanted to identify the different cellular states of trajectory. For this, I plotted expression of marker genes of cell types in the haematopoietic system on the UMAP plots. These included Ly6e (also known as Sca2) (haematopoietic stem cell marker), Fcgr3 (myeloid progenitor marker), Cd34

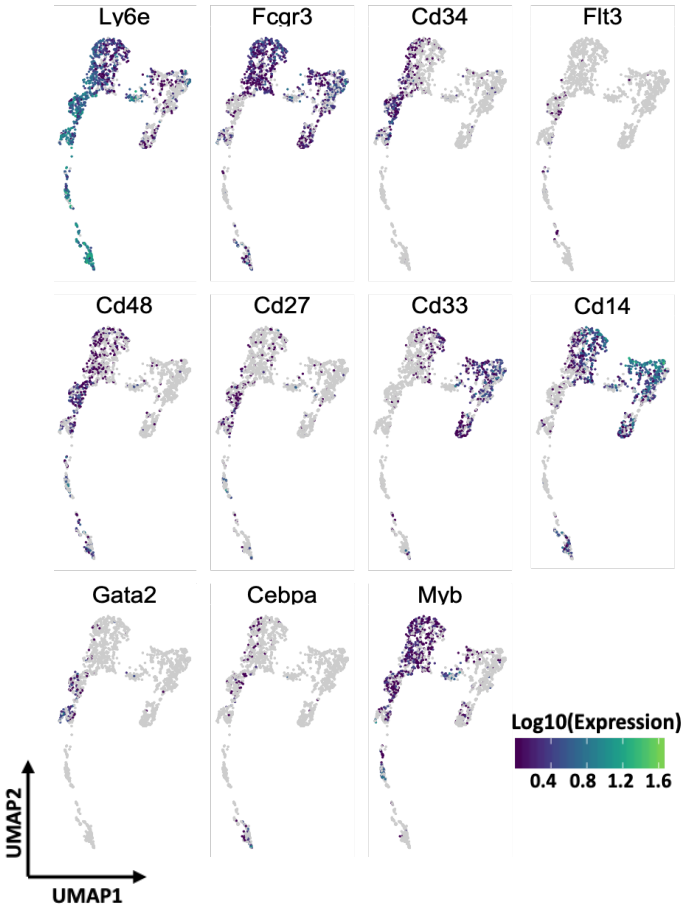
(haematopoietic stem cell marker), Flt3 (short term Haematopoietic Stem Cells (HSCs)/multipotent progenitor marker- also known as Cd135), Cd48 (lymphocyte marker), Cd27 (lymphocyte marker), Cd33 (myeloid marker), Cd14 (monocyte/macrophage marker), along with few transcription factors such as Gata2 (haematopoietic stem cell marker), Cebpα (myeloid progenitor marker) and Myb (myeloid progenitor marker) (Fig 6.2A).

Since all the cells subjected to scRNA-seq were selected for c-Kit positivity (early progenitor marker- also known as Cd117), a combination of c-Kit^{+}Ly6e^{hi}Cd34^{+}Flt3^{+} cells would represent a population of Lymphoid-primed multipotential like progenitors (LMPP). Indeed, I was able to distinguish an LMPP-like population in our UMAP plots (Fig 6.2B). These cells also had higher expression of Gata2 and were Myb positive, compatible with the LMPP-like phenotype (Fig 6.2A). Further, cells with Ly6e^{lo}Cd34^{+}Fcgr3^{+} expression highlighted a Granulocyte-Macrophage Progenitor (GMP) like population (Fig 6.2B). This population also had higher expression of Cebpα.

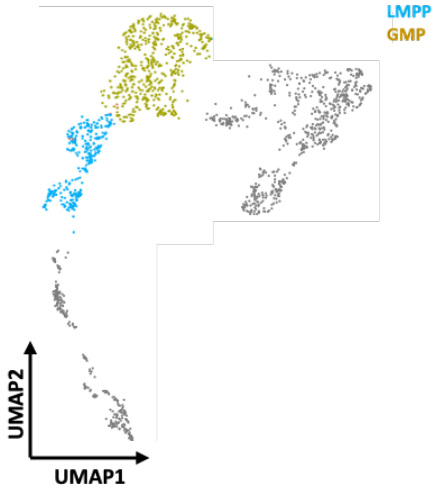
After characterizing these cell populations, I looked at the sample-wise composition of each population. Strikingly, 44.37% of the LMPP-like population was composed of *Kat2a* WT cells at 2 months post transplantation, whereas the rest of the samples had a contribution of ~15-20% (*Kat2a* WT cells 4 months- 15.89%, *Kat2a* NULL cells 2 months- 20.52%, *Kat2a* NULL cells 4 months- 19.20%). The higher percentage of LMPP-like population in the *Kat2a* WT cells at 2 months post transplantation was compatible with the experimental observation where loss of *Kat2a* accelerates *RUNX1-RUNX1T1(9a)* leukaemia progression.

Similarly, I looked at the GMP compartment, where 80.68% of the population was comprised of cells studied at 4 months from both *Kat2a* WT and *Kat2a* NULL genotypes (Fig 6.2D). This suggested that the cells collected at later time points follow a gradual leukaemic transformation trajectory. On the other hand, *Kat2a* WT cells and *Kat2a* NULL cells at the 2 months- time point together constituted 19.32% of the GMPs, where only 3.44% of the GMP population came from *Kat2a* NULL cells. This suggested that the *Kat2a* WT and NULL cells followed an abrupt leukaemia transformation trajectory, in line with the observations from genotype-specific and global single-cell trajectory analysis.

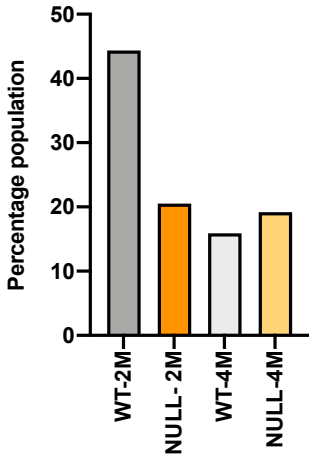
A


Ly6e
Fcgr3
Cd34
Flt3
Cd48
Cd27
Cd33
Cd14
Gata2
Cebpa
Myb
Log10(Expression)
0.4 0.8 1.2 1.6
UMAP2
UMAP1

B


LMPP
GMP
UMAP2
UMAP1

C


Percentage population
50
40
30
20
10
0
WT-2M
NULL- 2M
WT-4M
NULL-4M

D

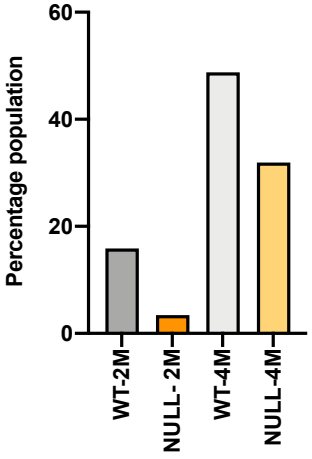
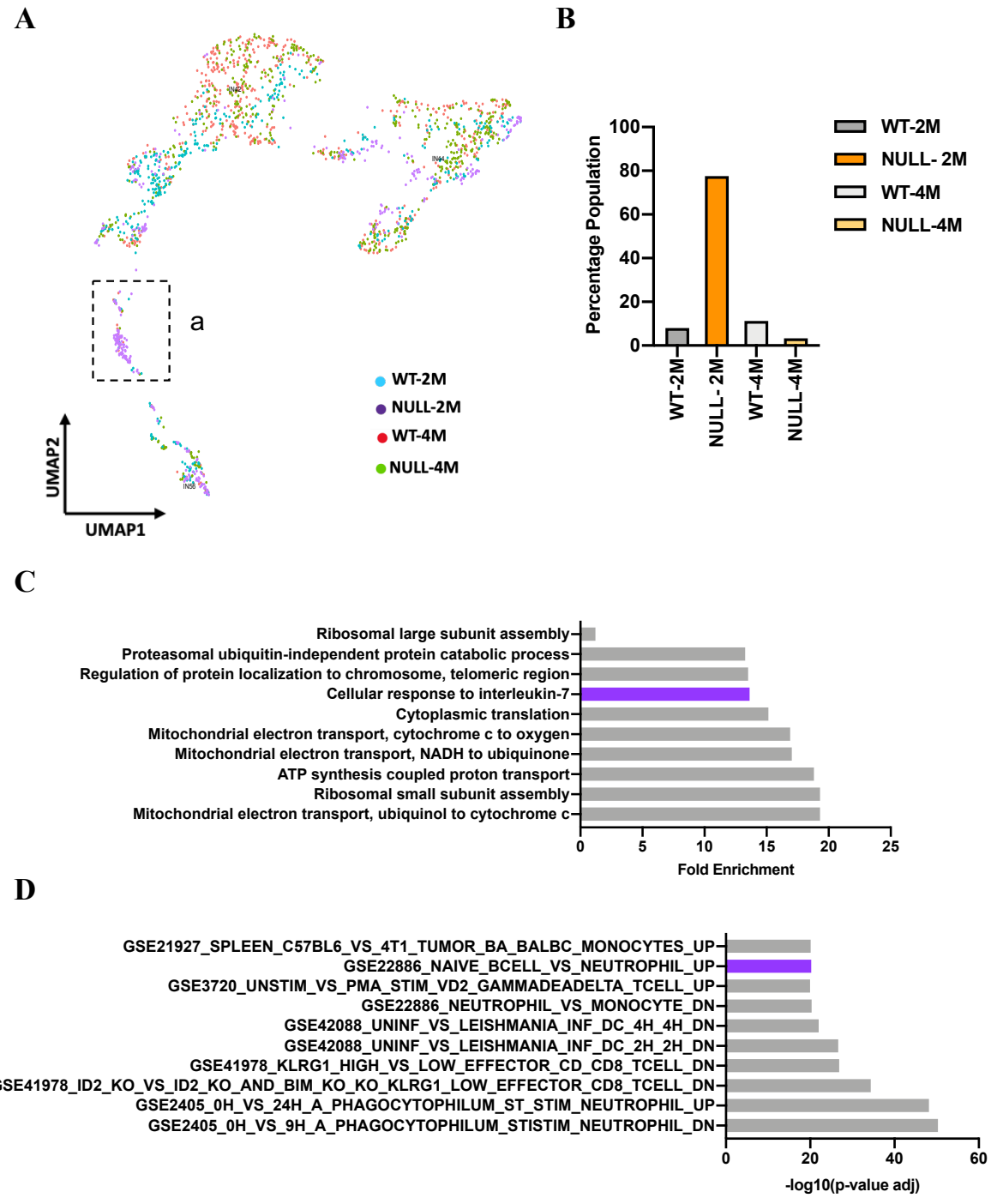

Percentage population
60
40
20
0
WT-2M
NULL- 2M
WT-4M
NULL-4M

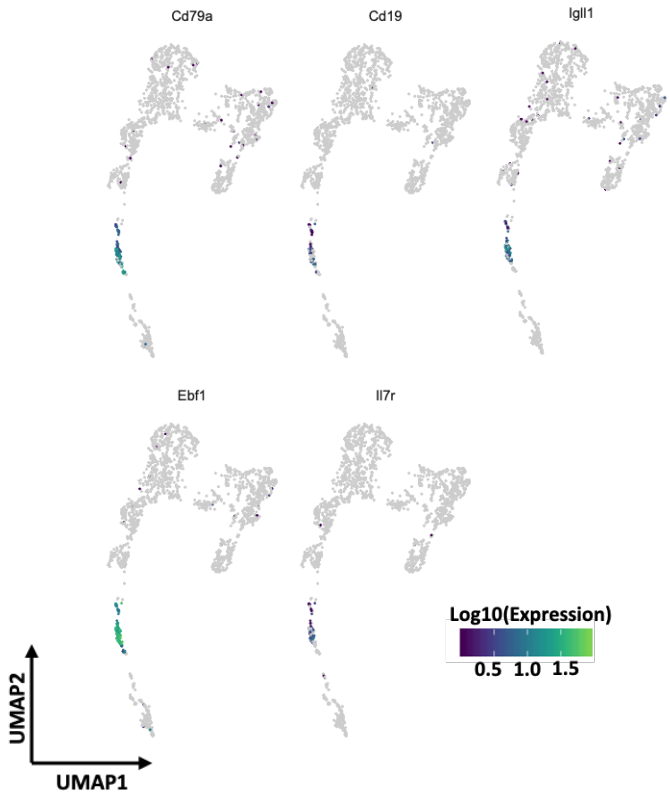
Figure 6.2: Representation of markers of haematopoietic hierarchy.
(A) Global UMAP representation of cells with expression of haematopoietic markers indicated, enabling identification of LMPP- c-Kit^{+}Ly6e^{+}CD34^{+}Flt3^{+}, and GMP- c-Kit^{+}Ly6e^{+}CD34^{+}Fcgr3^{+} cell populations. The colour represents average gene expression (log_{10} scale) where blue highlights cells with low expression and green represents high expression of the gene, **(B)** Global UMAP plot highlighting LMPP (blue) and GMP (green) population of cells, **(C)** Percentage composition of LMPP cell population by individual sample (*Kat2a* WT 2 months as WT-2M, *Kat2a* NULL 2 months as NULL-2M, *Kat2a* WT 4 months as WT-4M, *Kat2a* NULL 4 months as WT-4M), **(D)** Percentage composition of GMP cell population by individual sample.

6.3 Loss of *Kat2a* promotes differentiation towards B-cell lineage during pre-leukaemia transformation

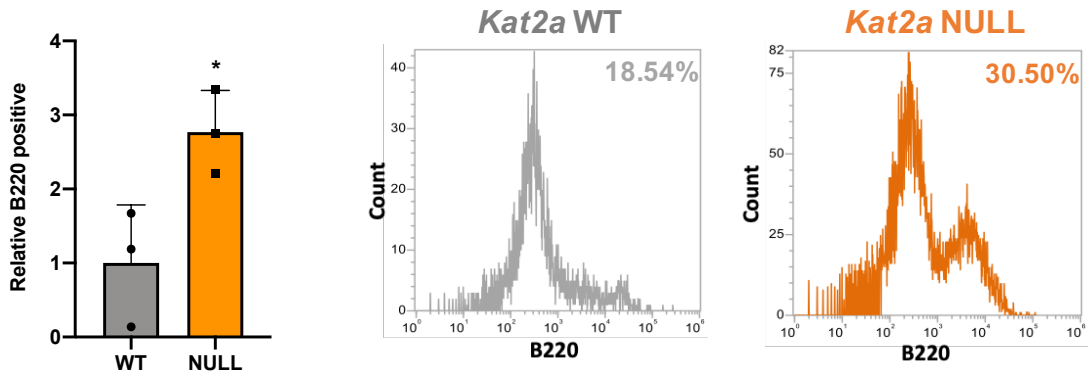
After characterizing the LMPP-like and GMP populations, I analysed the other arm of bifurcation observed from LMPP-like population of cells. This arm also captured *Kat2a* NULL cells exclusively from the 2 months time-point (for simplicity, this population is referred to as population 'a' going forward) (Fig 6.3A). 77.6% of the population 'a' comprised of *Kat2a* NULL cells at the 2 months time-point, indicative of an early consequence of *Kat2a* loss during leukaemia progression (Fig 6.3B). To identify the gene expression programmes characterizing population 'a', I conducted DESeq2 analysis (Love, Huber and Anders, 2014) to identify differentially expressed genes between population 'a' and the remaining cells, irrespective of genotype. I identified significant genes (p-adj<0.05, Bonferroni correction) which were upregulated in population 'a' and conducted Panther gene ontology analysis (H. Mi *et al.*, 2018) on this gene set (see Chapter-2 for methodology). This analysis highlighted pathways such as cellular response to Interleukin-7, which are specific to lymphocytes (Fig 6.3C). An overlap analysis with the Immunological gene expression signatures in MSigDB (Subramanian *et al.*, 2005; Liberzon *et al.*, 2015) also indicated a B-lymphocytes signature upregulated in population 'a' (Fig 6.3D). To further confirm this, I plotted average expression of B-cell surface markers, including Cd79a, Cd19, Igll1, Ebf1, and Il7r on the global UMAP plot (Fig 6.3D). Population 'a' specifically showed a higher average expression of these markers (Fig 6.3E), suggesting that loss of *Kat2a* promotes B-cell differentiation during early stages of *RUNX1-RUNX1T1(9a)* leukaemia.


A
a
WT-2M
NULL-2M
WT-4M
NULL-4M
UMAP2
UMAP1
B
Percentage Population
100
80
60
40
20
0
WT-2M
NULL- 2M
WT-4M
NULL-4M
C
Ribosomal large subunit assembly
Proteasomal ubiquitin-independent protein catabolic process
Regulation of protein localization to chromosome, telomeric region
Cellular response to interleukin-7
Cytoplasmic translation
Mitochondrial electron transport, cytochrome c to oxygen
Mitochondrial electron transport, NADH to ubiquinone
ATP synthesis coupled proton transport
Ribosomal small subunit assembly
Mitochondrial electron transport, ubiquinol to cytochrome c
0
5
10
15
20
25
Fold Enrichment
D
GSE21927_SPLEEN_C57BL6_VS_4T1_TUMOR_BA_BALBC_MONOCYTES_UP
GSE22886_NAIVE_BCELL_VS_NEUTROPHIL_UP
GSE3720_UNSTIM_VS_PMA_STIM_VD2_GAMMADEADELTA_TCELL_UP
GSE22886_NEUTROPHIL_VS_MONOCYTE_DN
GSE42088_UNINF_VS_LEISHMANIA_INF_DC_4H_4H_DN
GSE42088_UNINF_VS_LEISHMANIA_INF_DC_2H_2H_DN
GSE41978_KLRG1_HIGH_VS_LOW_EFFECTOR_CD_CD8_TCELL_DN
GSE41978_ID2_KO_VS_ID2_KO_AND_BIM_KO_KO_KLRG1_LOW_EFFECTOR_CD8_TCELL_DN
GSE2405_0H_VS_24H_A_PHAGOCYTOPHILUM_ST_STIM_NEUTROPHIL_UP
GSE2405_0H_VS_9H_A_PHAGOCYTOPHILUM_STISTIM_NEUTROPHIL_DN
0
20
40
60
-log10(p-value adj)

E


Cd79a
Cd19
Igll1
Ebf1
Il7r
Log10(Expression)
0.5 1.0 1.5
UMAP2
UMAP1

F


Relative B220 positive
WT
NULL
Kat2a WT
18.54%
Kat2a NULL
30.50%
Count
B220

G

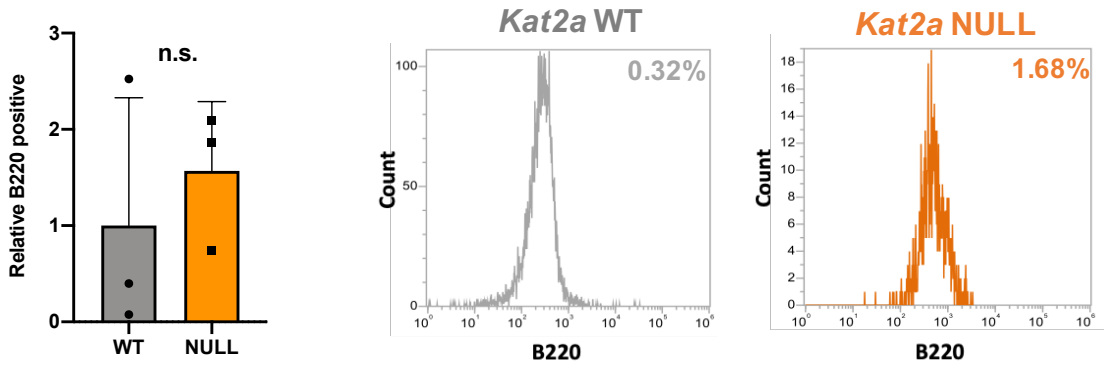

n.s.
Relative B220 positive
WT
NULL
Kat2a WT
0.32%
Kat2a NULL
1.68%
Count
B220

Figure 6.3: B-cell differentiation marker analysis.
(A) UMAP plot highlighting cluster *a* composed of *Kat2a* NULL cells at 2 months, **(B)** Sample-wise composition of cluster *a*, **(C)** Panther gene ontology analysis for genes upregulated in cluster *a*, where the pathway corresponding to B-cell differentiation is highlighted in purple, **(D)** Gene ontology analysis based on MSigDB overlap, highlighting a subset of genes involved in B-cell differentiation in purple, **(E)** UMAP plot with expression values of markers for B-cells. The log_{10} scaled expression represents average gene expression, where blue highlights cells with low expression and green represents high expression for that particular gene, **(F)** Flow cytometry analysis for B220, a B-cell marker, during *RUNX1-RUNX1T1(9a)* transformation *in vitro* at plate 2 (n=3/ sample, mean ± SD, Student's t-test, p= 0.039*), **(G)** Flow cytometry analysis for B220 during *RUNX1-RUNX1T1(9a)* transformation *in vitro* at plate 3 (n=3/ sample, mean ± SD, Student's t-test, p= 0.56).

To validate this observation functionally, I started with *RUNX1-RUNX1T1(9a) in vitro* transduced *Kat2a* WT and *Kat2a* NULL cells. These cells were maintained in the form of colony forming assay which allowed selection of transformants with each plating (Methods). With each plating, I studied the expression of B220, a B-cell marker individually in both *Kat2a* WT and *Kat2a* NULL cells using flow cytometry. I observed a significant increase in B220 expression in *Kat2a* NULL cells at plating 2 compared to *Kat2a* WT cells, confirming the scRNA-seq observation (Fig 6.3F). Interestingly, this increase in B220 expression was lost at plating 3 (Fig 6.3G), indicating that enhanced B-cell differentiation is consequential to early loss of *Kat2a.*

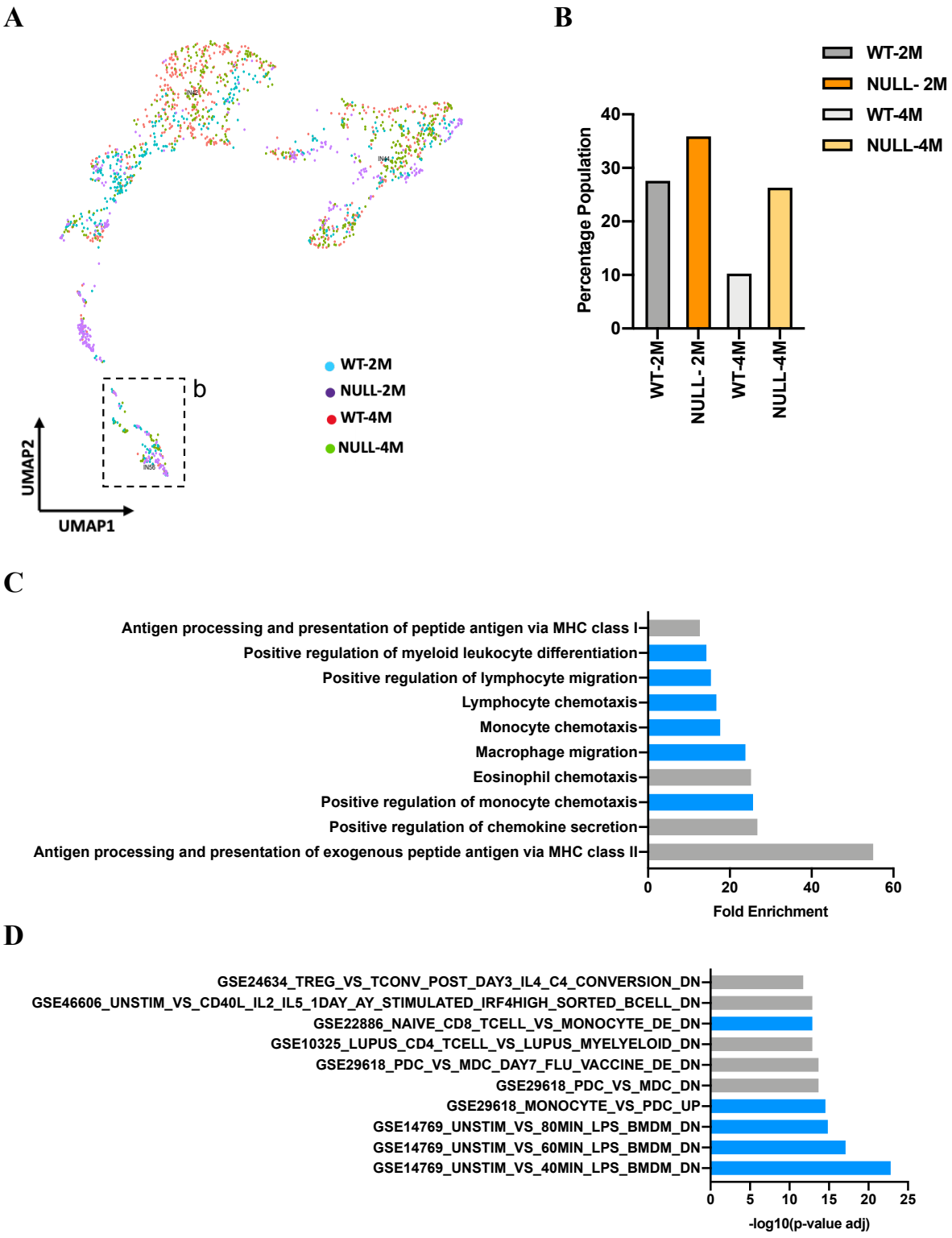
6.4 Loss of *Kat2a* promotes monocytic differentiation during pre-leukaemia transformation

After characterizing population 'a', I looked at a second group of cells, hereafter referred to as population 'b' (Fig 6.4A). This population had a higher percentage of *Kat2a* NULL cells (62.17%) compared to *Kat2a* WT cells (37.81%) (Fig 6.4B). To identify the gene expression programmes characterizing population 'b', I conducted DESeq2 analysis to identify differentially expressed genes between population 'b' and the remaining cells, irrespective of genotype. The gene ontology analysis for significantly upregulated genes (p-adj<0.05, Bonferroni correction) in population 'b' was done using Panther (Methods). This analysis

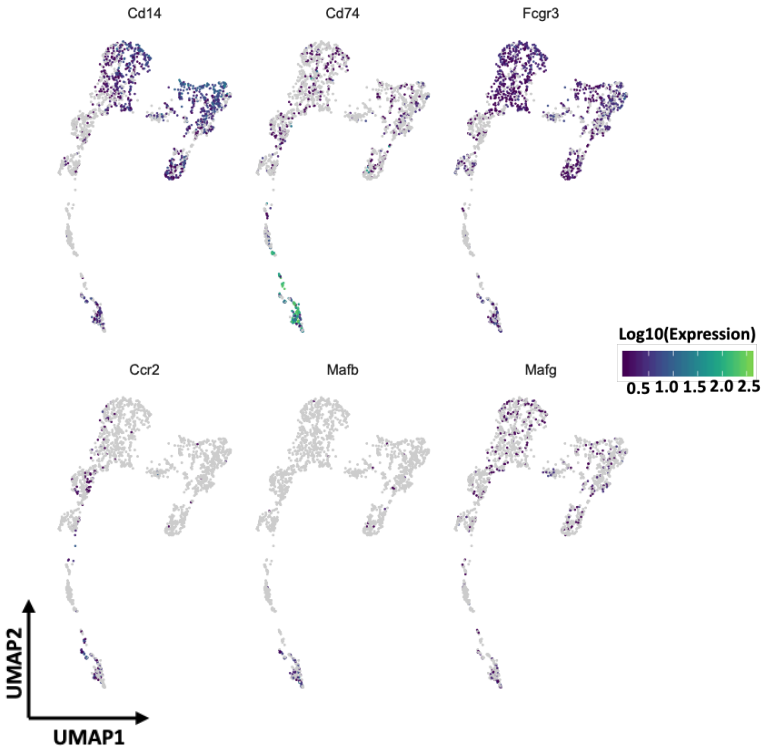
indicated an enrichment of genes involved in monocyte/macrophage development (Fig 6.4C). Similar observations were obtained upon performing an overlap with MSigDB Immunological Signatures database (methodology described previously) (Fig 6.4D). Cell surface markers associated with monocyte biogenesis including Cd14, Cd74, Fcgr3, Ccr2, Mafb, and Mafg displayed higher average expression in population 'b' compared to the rest of the cells (Fig 6.4E). Overall, the analysis highlighted differentiation of progenitor cells towards a monocytic lineage during the process of *RUNX1-RUNX1T1(9a)* leukaemia transformation.

Since scRNA-seq captured a higher proportion of *Kat2a* NULL cells differentiating to monocytes compared to *Kat2a* WT cells, I wanted to validate this observation in an experimental set-up. For this, I started with *RUNX1-RUNX1T1(9a) in vitro* transduced *Kat2a* WT and *Kat2a* NULL cells which were maintained in a colony forming assay (methodology described previously). Using flow cytometry analysis, I analysed the expression of F4/80, a monocyte-specific cell surface marker, at each plating. I observed a significant increase in F4/80 expression in *Kat2a* NULL cells at plating 2 compared to *Kat2a* WT cells, supporting the observation from scRNA-seq (Fig 6.4F). Similar to the observation for B-cell differentiation, there was no change in F4/80 expression at plating 3 (Fig 6.4G). These observations altogether validated findings of increased monocyte differentiation along with B-cell differentiation upon *Kat2a* loss during *RUNX1-RUNX1T1(9a)* leukaemia progression.

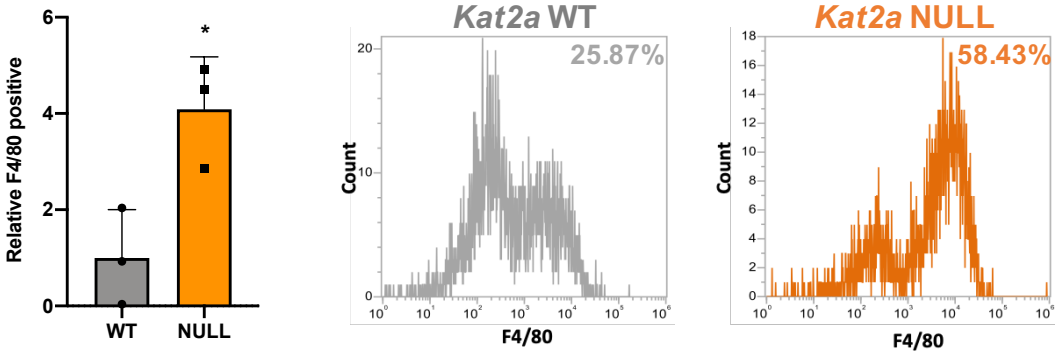
This is in line with my previous observations in the *MLL-AF9* leukaemia model, where loss of *Kat2a* promotes cellular differentiation. This was evident from a significant reduction in compact colony type along with an increase in mixed and dispersed colonies (Fig 6.4H). The differentiation observed upon *Kat2a* loss was towards a monocytic lineage based on *in vitro* analysis of primary *MLL-AF9* leukaemia cells (Fig 6.4I) (Domingues *et al.*, 2020). It was quite interesting to notice the monocytic differentiation associated with *Kat2a* loss in different models of AML, where in case of *RUNX1-RUNX1T1(9a)* loss of *Kat2a* accelerates the disease progression whereas in case of *MLL-AF9*, *Kat2a* loss leads to depletion of leukaemia stem like cells. The similarities observed in terms of monocytic differentiation indicates that loss of *Kat2a* enhances the probability of cell fate transitions, resulting in increased molecular heterogeneity and stochastic fate choices. This further suggests that *Kat2a* loss likely facilitates rather than determines fate acquisition, which consequently leads to its model-specific role.


A
WT-2M
NULL-2M
WT-4M
NULL-4M
b
UMAP2
UMAP1
B
WT-2M
NULL- 2M
WT-4M
NULL-4M
Percentage Population
0
10
20
30
40
WT-2M
NULL- 2M
WT-4M
NULL-4M
C
Antigen processing and presentation of peptide antigen via MHC class I
Positive regulation of myeloid leukocyte differentiation
Positive regulation of lymphocyte migration
Lymphocyte chemotaxis
Monocyte chemotaxis
Macrophage migration
Eosinophil chemotaxis
Positive regulation of monocyte chemotaxis
Positive regulation of chemokine secretion
Antigen processing and presentation of exogenous peptide antigen via MHC class II
0
20
40
60
Fold Enrichment
D
GSE24634_TREG_VS_TCONV_POST_DAY3_IL4_C4_CONVERSION_DN
GSE46606_UNSTIM_VS_CD40L_IL2_IL5_1DAY_AY_STIMULATED_IRF4HIGH_SORTED_BCELL_DN
GSE22886_NAIVE_CD8_TCELL_VS_MONOCYTE_DE_DN
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_MYELYELOID_DN
GSE29618_PDC_VS_MDC_DAY7_FLU_VACCINE_DE_DN
GSE29618_PDC_VS_MDC_DN
GSE29618_MONOCYTE_VS_PDC_UP
GSE14769_UNSTIM_VS_80MIN_LPS_BMDM_DN
GSE14769_UNSTIM_VS_60MIN_LPS_BMDM_DN
GSE14769_UNSTIM_VS_40MIN_LPS_BMDM_DN
0
5
10
15
20
25
-log10(p-value adj)

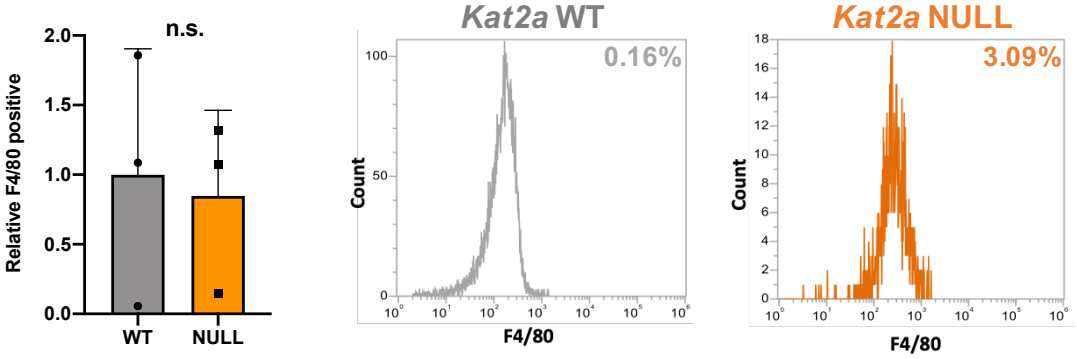
E


Cd14
Cd74
Fcgr3
Ccr2
Mafb
Mafg
Log10(Expression)
0.5 1.0 1.5 2.0 2.5
UMAP2
UMAP1

F


Relative F4/80 positive
WT
NULL
*
Kat2a WT
25.87%
Kat2a NULL
58.43%
Count
F4/80

G


Relative F4/80 positive
n.s.
WT
NULL
Kat2a WT
0.16%
Kat2a NULL
3.09%
Count
F4/80

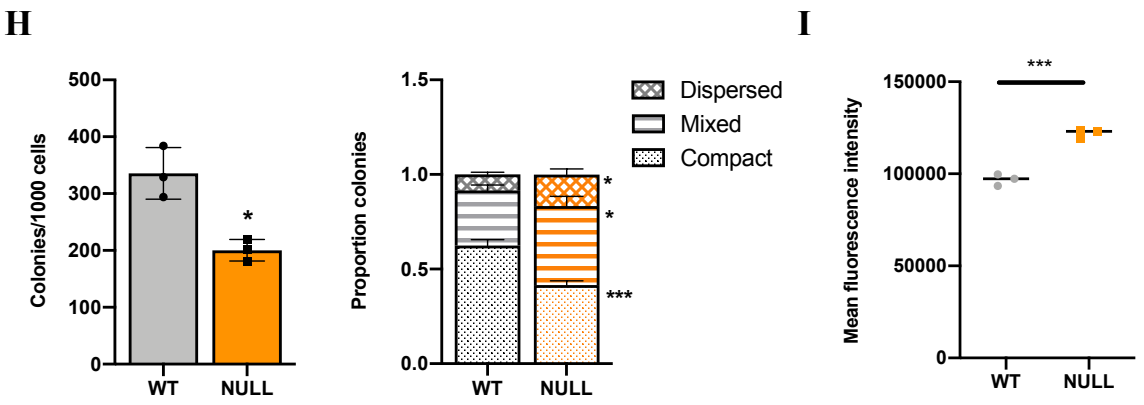


Figure 6.4: Monocyte marker analysis.

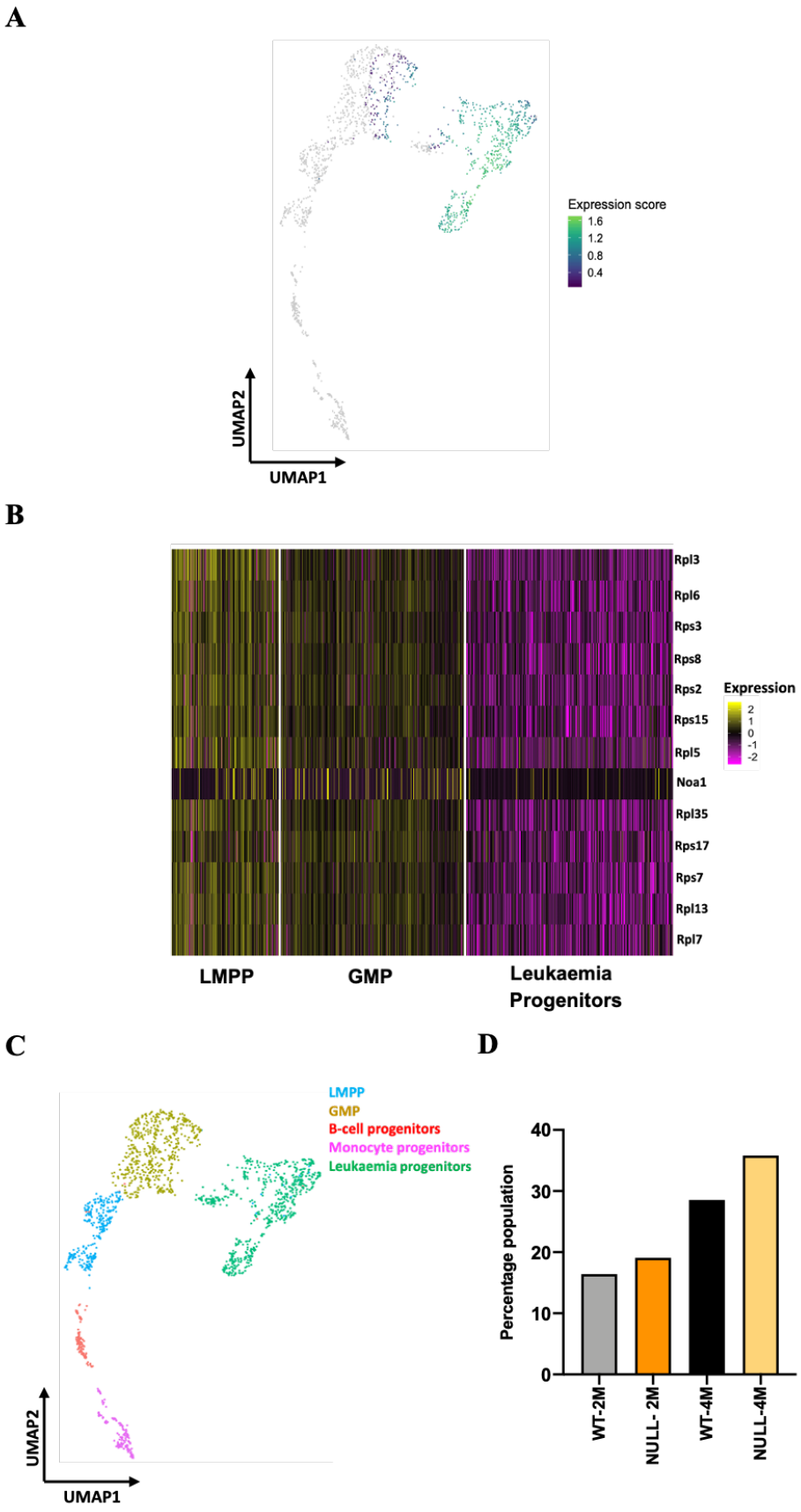
(A) UMAP plot highlighting cluster *b* composed of *Kat2a* NULL cells at 2 months and 4 months, **(B)** Sample-wise composition of cluster *b*, **(C)** Panther gene ontology analysis for genes upregulated in cluster *b*, where the pathway corresponding to monocytic differentiation is highlighted in blue, **(D)** Gene ontology analysis based on MSigDB overlap highlighting the set of genes corresponding to monocyte differentiation in blue, **(E)** UMAP plot of cells coloured by average expression of monocyte marker genes. The log_{10} scaled expression represents average gene expression where blue highlights cells with low expression and green represents high expression for that particular gene, **(F)** Flow cytometry analysis for F4/80, a monocyte marker, during *RUNX1-RUNX1T1(9a)* transformation *in vitro* at plate 2 (n=3/ sample, mean ± SD, Student's t-test, p= 0.0229*), **(G)** Flow cytometry analysis for F4/80 during *RUNX1-RUNX1T1(9a)* transformation *in vitro* at plate 3 (n=3/ sample, mean ± SD, Student's t-test, p= 0.8221), **(H)** Colony forming assay for *in vitro MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL cells (left) (n=3/ sample, mean ± SD, Student's t-test, p= 0.0225*); distribution of colonies highlighting proportion of compact (n=3/ sample, mean ± SD, Student's t-test, p= 0.001***), mixed (n=3/ sample, mean ± SD, Student's t-test, p= 0.0292*) and dispersed (n=3/ sample, mean ± SD, Student's t-test, p= 0.0264*) colonies (right), **(I)** Mean fluorescence intensity of Mac1, a monocyte marker, upon flow cytometry analysis of *in vitro MLL-AF9* transformed *Kat2a* WT and *Kat2a* NULL cells (n=3/ sample, mean ± SD, Student's t-test, p= 0.0005*).

6.5 *Kat2a* loss aids in accumulation of transformed pre-leukaemic cells

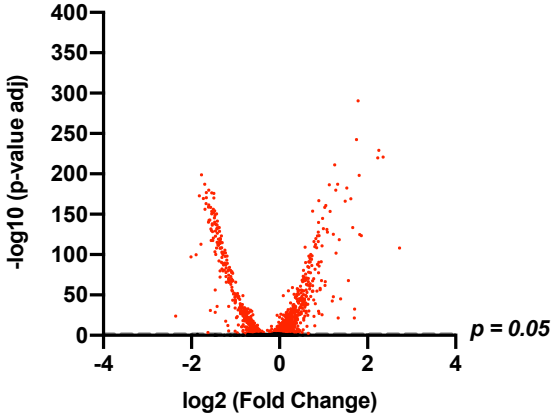
After identifying the cell populations in the *RUNX1-RUNX1T1(9a)* leukaemia transformation trajectory, including LMPP-candidate, GMPs, B-cell and monocytes, I wanted to characterize the population of cells which were not part of the global trajectory (Fig 6.1E, highlighted in grey). Since the myeloid arm of the trajectory followed a haematopoietic hierarchy from LMPP

candidate to GMP, I hypothesized that the population of cells not forming a trajectory could be *RUNX1-RUNX1T1(9a)* transformed leukaemia progenitor cells. For this, I looked at direct target genes of *RUNX1-RUNX1T1* based on ChIP-seq analysis in a previous study (Ptasinska *et al.*, 2012), and plotted their average expression on the global UMAP plot (Fig 6.5A). Interestingly, *RUNX1-RUNX1T1* targets had a higher average expression in this compartment compared to the remaining cells in the trajectory, suggestive of leukaemia-initiating events. To confirm this, I also looked at *Kat2a* targets based on unpublished ChIP-seq data in Pina lab. The ChIP-seq experiment was done on Kasumi-1 cells, which are representative of *RUNX1-RUNX1T1* leukaemia. *Kat2a* target genes were enriched in gene expression programmes related to cytoplasmic translation and ribosome biogenesis, compatible with the observation from scRNA-seq data (Chapter-4). These *Kat2a* target genes were highly expressed in LMPP-like population of cells, with a gradual reduction in GMPs, and a further reduction in the 'Leukaemia Progenitors' compartment (Fig 6.5B). Further, I looked at the composition of the leukaemia progenitor compartment of cells (Fig 6.5C) and found that it was predominantly populated with the *Kat2a* WT and *Kat2a* NULL cells at 4 months post-transplantation (Fig 6.5D). The gradual accumulation of leukaemia progenitors with time coincided with the progressive *RUNX1-RUNX1T1(9a)* transformation (Fig 6.5D). The increased accumulation of leukaemia progenitors upon *Kat2a* loss at respective time points was in-line with the *in vivo* observation of accelerated leukaemia progression upon loss of *Kat2a*.

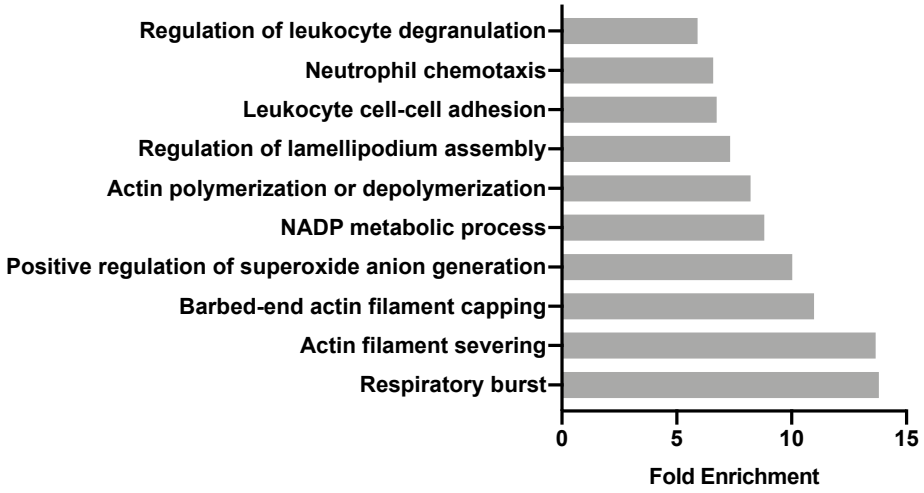
To further characterize this population of cells, differential gene expression analysis was performed using DESeq2, comparing leukaemia progenitor cells against other cells (Fig 6.5E). There were 1877 genes (p-adj<0.05, Bonferroni correction) which were differentially expressed in leukaemia progenitors compared to the rest of the population. Out of 1877 genes, 897 genes were upregulated, whereas 980 genes were downregulated (Fig 6.5E). These upregulated genes were mostly associated with general gene expression categories such as respiratory burst and haematopoietic hierarchy, including leukocyte maintenance and differentiation (Fig 6.5F and 6.5G). On the other hand, the downregulated genes captured gene signatures associated with *Kat2a* loss, namely, mitochondrial ATP transport, translation/ribosome biogenesis (Fig. 6H), and lymphocyte maintenance (Fig. 6I), which was compatible with the observation of *Kat2a* mediated accumulation of leukaemia progenitors.


A
Expression score
1.6
1.2
0.8
0.4
UMAP2
UMAP1
B
Rpl3
Rpl6
Rps3
Rps8
Rps2
Rps15
Rpl5
Noa1
Rpl35
Rps17
Rps7
Rpl13
Rpl7
Expression
2
1
0
-1
-2
LMPP
GMP
Leukaemia Progenitors
C
LMPP
GMP
B-cell progenitors
Monocyte progenitors
Leukaemia progenitors
UMAP2
UMAP1
D
Percentage population
40
30
20
10
0
WT-2M
NULL- 2M
WT-4M
NULL-4M
NULL-4M
WT-4M
NULL- 2M
WT-2M

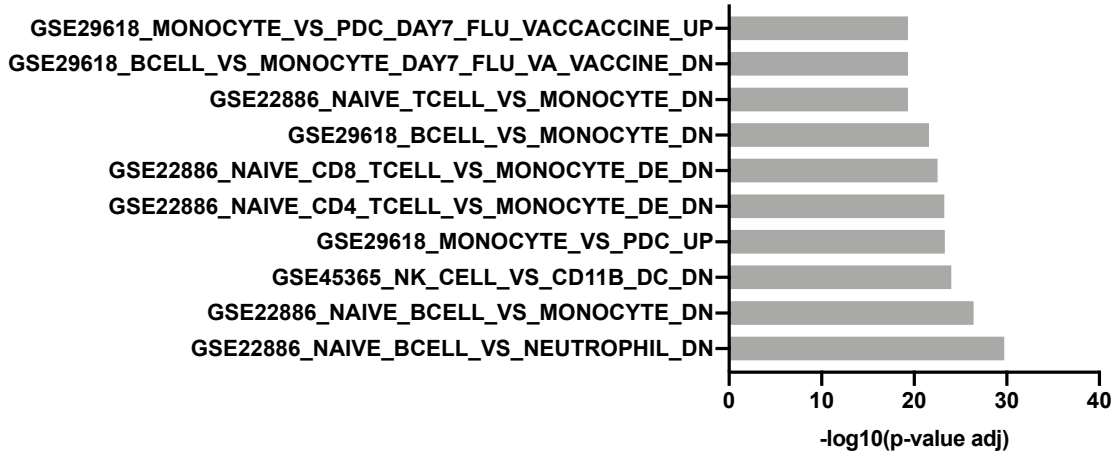
E


400
350
300
250
200
150
100
50
0
-log10 (p-value adj)
p = 0.05
-4
-2
0
2
4
log2 (Fold Change)

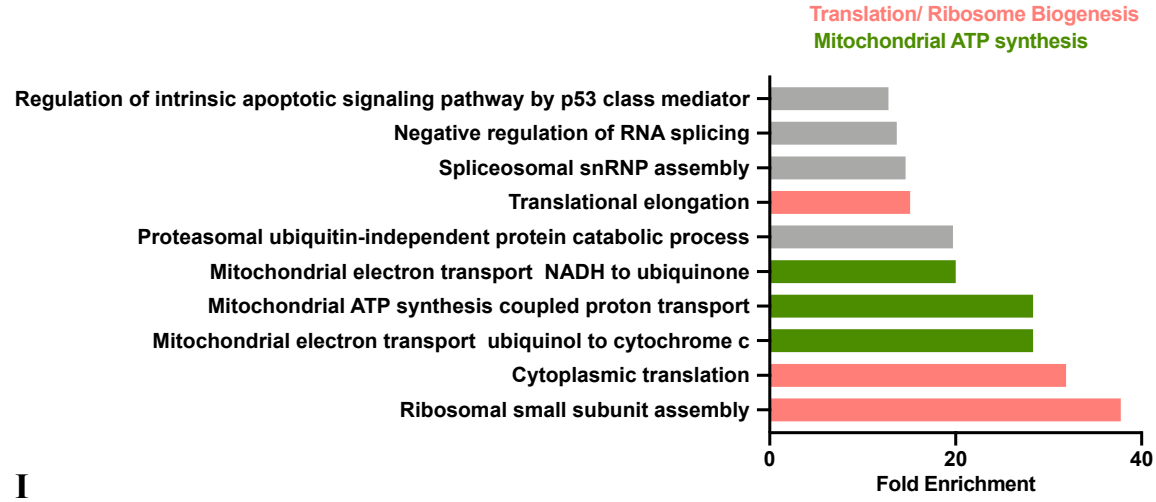
F


Regulation of leukocyte degranulation
Neutrophil chemotaxis
Leukocyte cell-cell adhesion
Regulation of lamellipodium assembly
Actin polymerization or depolymerization
NADP metabolic process
Positive regulation of superoxide anion generation
Barbed-end actin filament capping
Actin filament severing
Respiratory burst
0
5
10
15
Fold Enrichment

G


GSE29618_MONOCYTE_VS_PDC_DAY7_FLU_VACCACCINE_UP
GSE29618_BCELL_VS_MONOCYTE_DAY7_FLU_VA_VACCINE_DN
GSE22886_NAIVE_TCELL_VS_MONOCYTE_DN
GSE29618_BCELL_VS_MONOCYTE_DN
GSE22886_NAIVE_CD8_TCELL_VS_MONOCYTE_DE_DN
GSE22886_NAIVE_CD4_TCELL_VS_MONOCYTE_DE_DN
GSE29618_MONOCYTE_VS_PDC_UP
GSE45365_NK_CELL_VS_CD11B_DC_DN
GSE22886_NAIVE_BCELL_VS_MONOCYTE_DN
GSE22886_NAIVE_BCELL_VS_NEUTROPHIL_DN
0
10
20
30
40
-log10(p-value adj)

H


Translation/ Ribosome Biogenesis
Mitochondrial ATP synthesis
Regulation of intrinsic apoptotic signaling pathway by p53 class mediator
Negative regulation of RNA splicing
Spliceosomal snRNP assembly
Translational elongation
Proteasomal ubiquitin-independent protein catabolic process
Mitochondrial electron transport NADH to ubiquinone
Mitochondrial ATP synthesis coupled proton transport
Mitochondrial electron transport ubiquinol to cytochrome c
Cytoplasmic translation
Ribosomal small subunit assembly
0
20
40
Fold Enrichment

I

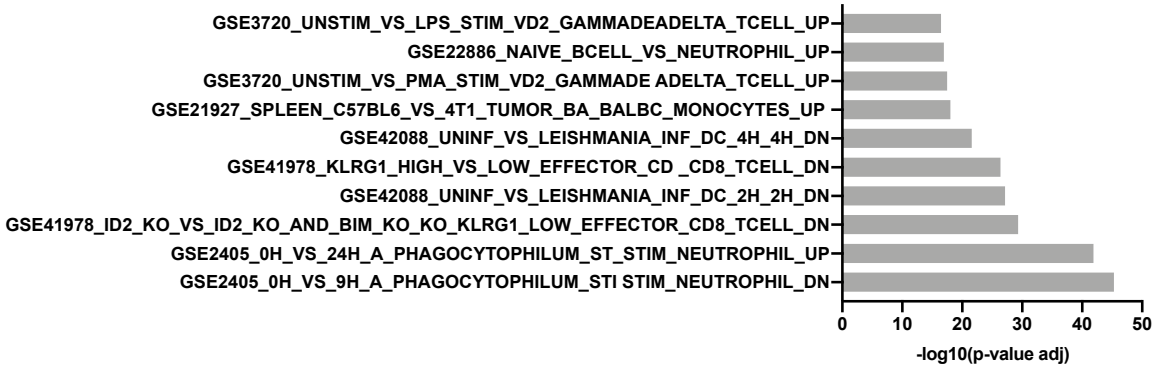

GSE3720_UNSTIM_VS_LPS_STIM_VD2_GAMMADEADELTA_TCELL_UP
GSE22886_NAIVE_BCELL_VS_NEUTROPHIL_UP
GSE3720_UNSTIM_VS_PMA_STIM_VD2_GAMMADE ADELTA_TCELL_UP
GSE21927_SPLEEN_C57BL6_VS_4T1_TUMOR_BA_BALBC_MONOCYTES_UP
GSE42088_UNINF_VS_LEISHMANIA_INF_DC_4H_4H_DN
GSE41978_KLRG1_HIGH_VS_LOW_EFFECTOR_CD _CD8_TCELL_DN
GSE42088_UNINF_VS_LEISHMANIA_INF_DC_2H_2H_DN
GSE41978_ID2_KO_VS_ID2_KO_AND_BIM_KO_KO_KLRG1_LOW_EFFECTOR_CD8_TCELL_DN
GSE2405_0H_VS_24H_A_PHAGOCYTOPHILUM_ST_STIM_NEUTROPHIL_UP
GSE2405_0H_VS_9H_A_PHAGOCYTOPHILUM_STI STIM_NEUTROPHIL_DN
0
10
20
30
40
50
-log10(p-value adj)

Figure 6.5: Characterization of leukaemia compartment.

(A) UMAP plot of cells coloured by average expression of *RUNX1-RUNX1T1* target genes obtained from a previous study (Ptasinska *et al.*, 2012), **(B)** Heatmap of gene expression values in different cell compartments for *Kat2a* direct targets identified in an unpublished ChIP-seq dataset from MB-3 treated Kasumi-1 cells available in the lab. Downregulation of genes is highlighted in pink, **(C)** UMAP plot of cells coloured by identified compartments; LMPP (blue), GMP (golden), B-cell progenitors (red), Monocyte progenitors (pink), and the leukaemia compartment (green), **(D)** Composition of leukaemia progenitors cells by individual sample (*Kat2a* WT 2 months as WT-2M, *Kat2a* NULL 2 months as NULL-2M, *Kat2a* WT 4 months as WT-4M, *Kat2a* NULL 4 months as WT-4M), **(E)** Volcano plot representing differentially expressed genes in leukaemia compartment (red) with p-adj<0.05*, **(F)** Gene Ontology analysis using Panther 14.0 for upregulated genes in leukaemia compartment obtained from DESeq2 analysis (p<0.05*), **(G)** Gene Ontology analysis based on MSigDB overlap for upregulated genes in leukaemia compartment (FDR<0.05), **(H)** Gene Ontology analysis using Panther 14.0 for downregulated genes in leukaemia compartment highlighting translation/ribosome biogenesis (pink) and mitochondrial translation (green) (p<0.05*), **(I)** Gene Ontology analysis based on MSigDB overlap for downregulated genes in leukaemia compartment (FDR<0.05).

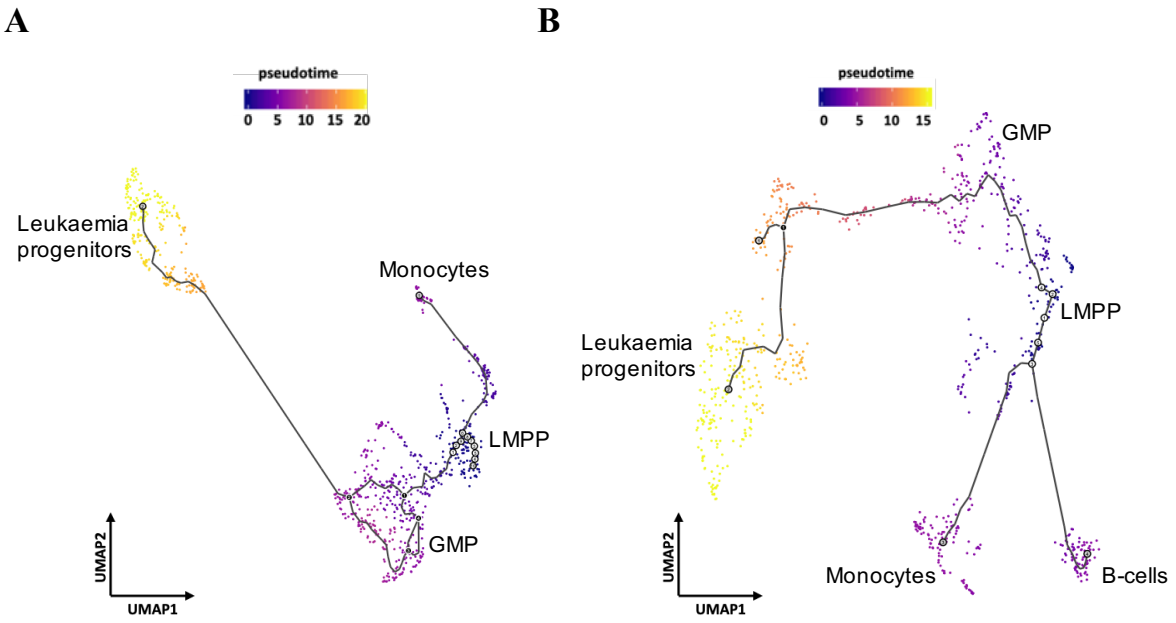


Figure 6.6: Pseudotime trajectory plots highlighting different population of cells.
Pseudotime trajectory representation on a UMAP plot for **(A)** *Kat2a* WT cells, **(B)** *Kat2a* NULL cells, including cells collected at both 2 months and 4 months post transplantation. Populations of cells identified based on cell surface markers**,** LMPP- c-Kit^{+}Ly6e^{+}CD34^{+}Flt3^{+}, GMP- c-Kit^{+}Ly6e^{+}CD34^{+}Fcgr3^{+}, Monocytes- Cd14^{+}Cd74^{+}Mafb^{+}, B-cells- Cd79a^{+}Cd19^{+}Il7r^{+}, Leukaemia progenitors- c-Kit^{+}FcγR^{+}RUNX1-RUNX1T1^{+} are highlighted.

Overall, the characterization of different cellular compartments of the trajectory highlighted the overall pre-leukaemia transformation process starting from LMPP-like, progressing to GMP, and further leading to the accumulation of leukaemia progenitors (Fig 6.6).

6.6 Loss of *Kat2a* increases transcriptional variability during pre-leukaemia progression

After studying the role of *Kat2a* in promoting cellular variability during leukaemia initiation and given its central role in limiting cell-to-cell transcriptional variability, I wanted to interrogate a potential link between loss of *Kat2a*, its consequent increase in transcriptional variability, and pre-leukaemia progression. Briefly, as mentioned in Chapter-1, our lab has recently deciphered the role of *Kat2a* in *MLL-AF9* leukaemia using a conditional *Kat2a* knockout mice model, where we observed that loss of *Kat2a* impacts the long-term preservation of functional leukaemia stem-like cells (LSC) (Domingues *et al.*, 2020). Upon performing scRNA-seq of *Kat2a* WT and NULL leukaemia, an increase in cell-to-cell transcriptional variability was observed in *Kat2a* NULL cells, suggestive of a role for *Kat2a* in promoting transcriptional stability and altogether, revealing cell-to-cell transcriptional variability as a mechanism where leukaemia transformed cells lose the capability to maintain a balance between self-renewal and differentiation.

In order to study this in the context of pre-leukaemia, I first compared cell-to-cell transcriptional variability at a global scale upon *Kat2a* loss. Transcriptional variability was calculated based on distance to median (DM), which was used to identify the top 500 most variable genes in each sample. After defining the top 500 variable genes in each sample, a

pairwise correlation method was employed which calculated cell-to-cell variability based on a Spearman's correlation coefficient (Mohammed *et al.*, 2017) (Methods). The global analysis indicated an increase in transcriptional variability upon loss of *Kat2a* at respective time points (Fig 6.7A) in-line with previous observations in the lab (Domingues *et al.*, 2020). The increase in transcriptional variability upon *Kat2a* loss was more evident at 2 months compared to 4 months, compatible with the cell fate diversification observed at early stages of leukaemia transformation. Further, there was a reduction in transcriptional variability at 4 months compared to 2 months in both the genotypes, suggesting that increase in cell-to-cell transcriptional variability is a hallmark of early pre-leukaemia transformation event. However, it is worth noting that enhanced transcriptional instability consequent to a loss of *Kat2a* may be a reflection of the underlying increase in cellular diversity at the 2 months time point.

After making global comparisons, I wanted to understand whether different compartments of cells showed differences in cell-to-cell transcriptional variability. For this, I compared the LMPP-like, GMP, and leukaemia progenitor cell populations using Spearman's correlation. Compared to LMPP-candidate, GMPs showed a reduction in transcriptional variability, compatible with the cellular fate trajectory observed at early time points of *RUNX1-RUNX1T1(9a)* leukaemia initiation (Fig 6.7B). As the haematopoietic hierarchy progresses, cells become relatively more committed, and hence, a reduction in transcriptional variability was observed in case of GMPs. However, leukaemia progenitor cells showed an increase in transcriptional variability, potentially due to the presence of leukaemia-initiating events leading to leukaemia establishment (Fig 6.7B). This was in-line with the global comparisons which confirmed the role of cell-to-cell transcriptional variability during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation.

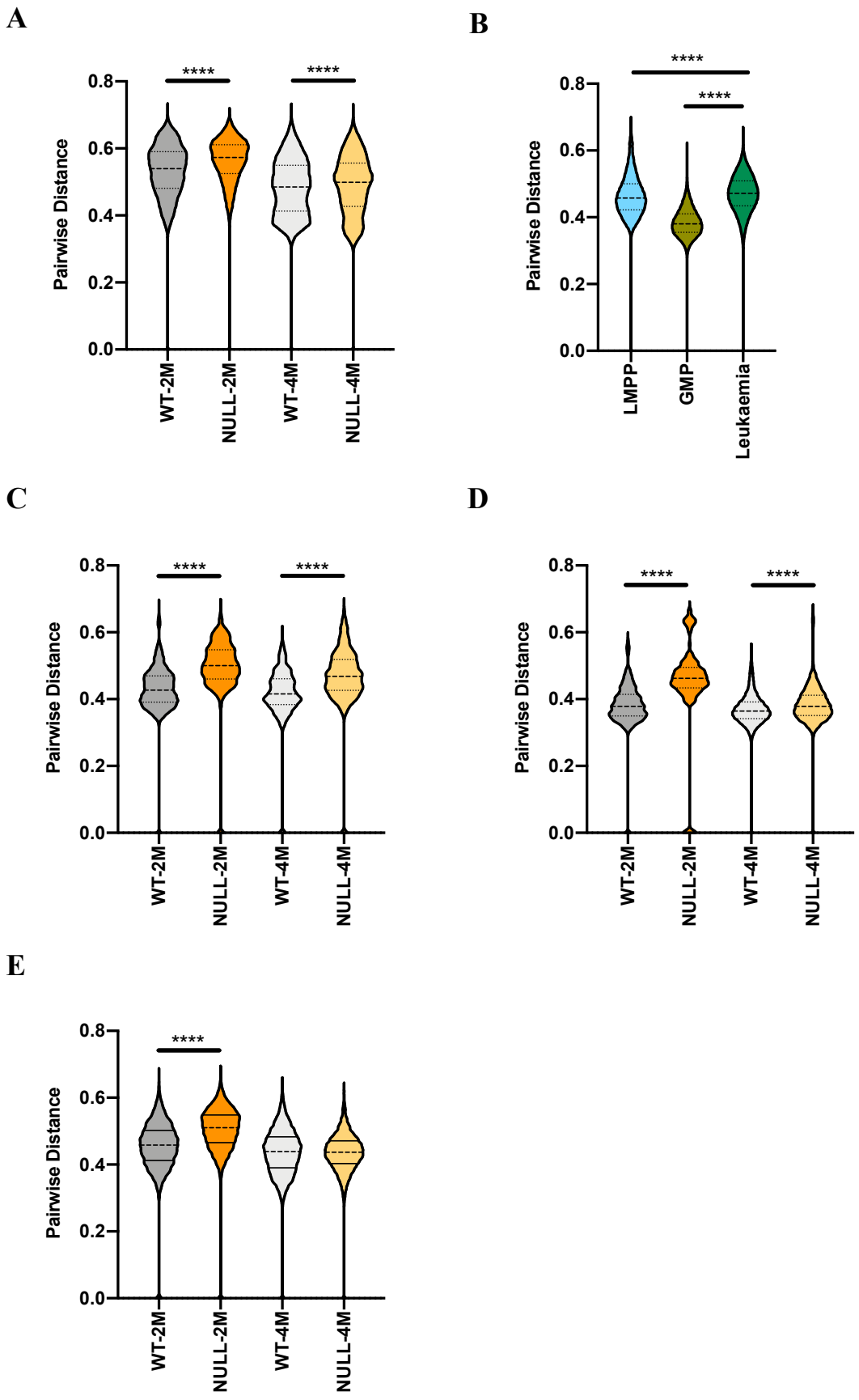

A
B
C
D
E
Pairwise Distance
0.0
0.2
0.4
0.6
0.8
****
WT-2M
NULL-2M
WT-4M
NULL-4M
LMPP
GMP
Leukaemia

Figure 6.7: Pairwise correlation measure for transcriptional variability.
(A) Violin plot for global transcriptional variability based on pairwise correlation method (p<0.0001**** *Kat2a* WT 2 months vs *Kat2a* NULL 2 months, p<0.0001**** *Kat2a* WT 4 months vs *Kat2a* NULL 4 months), **(B)** Violin plot representing transcriptional variability based on pairwise correlation method for different cell compartments (p<0.0001**** LMPP vs leukaemia, p<0.0001**** GMP vs leukaemia), **(C)** Violin plot for transcriptional variability based on pairwise correlation method for samples within LMPP compartment (p<0.0001**** *Kat2a* WT 2 months vs *Kat2a* NULL 2 months, p<0.0001**** *Kat2a* WT 4 months vs *Kat2a* NULL 4 months), **(D)** Violin plot for transcriptional variability based on pairwise correlation method for samples within GMP compartment (p<0.0001**** *Kat2a* WT 2 months vs *Kat2a* NULL 2 months, p<0.0001**** *Kat2a* WT 4 months vs *Kat2a* NULL 4 months), **(E)** Violin plot for transcriptional variability based on pairwise correlation method for samples within leukaemia compartment (p<0.0001**** *Kat2a* WT 2 months vs *Kat2a* NULL 2 months).

Further, I looked at the cell-to-cell transcriptional variability for individual samples within each cell compartment. In LMPP candidates, there was an increase in transcriptional variability in *Kat2a* NULL cells compared to *Kat2a* WT cells, which was more prominent at 2 months (Fig 6.7C). Overall, a reduction in cell-to-cell transcription variability was observed from 2 months to 4 months, compatible with the global observation for each genotype. I extended the analysis for GMPs, where I observed a similar reduction in cell-to-cell transcriptional variability (Fig 6.7). Again, the genotype-specific differences were more prominent at 2 months compared to 4 months. Finally, in case of leukaemia progenitors, there was again an increase in cell-to-cell variability at 2 months upon *Kat2a* loss, followed by a reduction between 2 months and 4 months (Fig 6.7). However, no differences in cell-to-cell variability were observed in *Kat2a* NULL cells at 4 months compared to *Kat2a* WT, suggesting that leukaemia is likely established by that time point. This supports the role of transcriptional variability in contributing to *RUNX1-RUNX1T1(9a)* leukaemia initiation. However, it is worth noting that the reduction in cell-to-cell transcriptional variability at 4 months-time point could be a technical artifact owing to the small size of individual cell populations.

6.7 Inhibition of KAT2A rearranges chromatin accessibility pattern in Kasumi-1 cells

Having studied the role of cell-to-cell transcriptional variability associated with loss of *Kat2a* in promoting cellular diversity and accumulation of leukaemia progenitor cells during *RUNX1-RUNX1T1(9a)* pre-leukaemia, I wanted to understand the underlying causes of increased transcriptional variability during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation. Recent evidence suggests that transcriptional variability is consequential to inherent epigenetic variability and contributes towards disease relapse and diagnosis, (Landau *et al.*, 2014; Pan *et al.*, 2015; Li *et al.*, 2016). Therefore, I interrogated whether this gain in transcriptional variability is associated with underlying epigenetic variability. For this, I analysed a pilot study conducted in Pina lab on Kasumi-1 cells, a human cell line representative of *RUNX1-RUNX1T1(9a)* leukaemia where inhibition of KAT2A activity was achieved by using MB-3 as an inhibitor (Methods). These cells were then subjected to single-cell ATAC sequencing (scATAC-seq), a methodology to study variation in chromatin accessibility at single-cell level, based on Tn5 transposase activity (Buenrostro *et al.*, 2015b). The library preparation was done by Pina lab members prior to my lab joining, using C_1 Single-Cell Auto Prep System (Fluidigm, Inc.) (Methods).

All generated libraries were subjected to paired-end sequencing. Adapter sequences were trimmed from FastQ files and the paired-end reads were aligned to the hg19 reference genome using Bowtie2 (Langmead and Salzberg, 2012). Duplicate reads were removed using Picard Tools (http://picard.sourceforge.net) and peak calling was done using MACS2 (Zhang *et al.*, 2008). Following peak calling, I obtained 74,284 peaks from control DMSO cells and 111, 898 peaks from MB-3 treated cells (Fig 6.8). The higher number of peaks obtained in MB-3 treated cells could be a consequence of more cells subjected to sequencing compared to DMSO (38 cells for DMSO and 50 cells for MB-3). To generate a combined matrix representing peaks from both the samples, these two peak sets were merged using Bedtools 2.27.0 (Quinlan and Hall, 2010). This generated a combined matrix with peak information from both DMSO and MB-3 with a total of 172,299 peaks. I used stringent filtering criteria for further analysis in order to exclude technical noise arising from single-cell data and due to the unequal number of cells sequenced for each sample. In order to assess reliable peaks, each peak was considered as

a confident peak only when it is present in 15 out of 88 cells in total. The rest of the peaks were filtered out for downstream analysis. This strict filtering led to a total of 4,157 peaks which were used in downstream analysis. Post filtering, the peaks were subjected to differential accessibility analysis based on Fisher's exact test and information gain (Methods). From this analysis, I identified significant differentially accessible peaks (p-adj <0.05, Bonferroni and Benjamin Hochberg correction). I obtained 50 peaks which were characterized as DMSO unique, 520 peaks which were characterized as MB-3 unique, and 3,587 peaks which were attributed as common peaks (Fig 6.8). Within the common peak subset, there were no significant differences in terms of differential accessibility upon inhibition of KAT2A.

After identifying the unique and common subsets of peaks and confirming no differences within common subset, I looked at the peak accessibility frequency of the unique peak subsets (Fig 6.9A). The DMSO unique peaks showed an average peak frequency of 41.73% in DMSO cells, whereas MB-3 unique peaks had an average peak frequency of 29.96% in MB-3 cells, slightly lower than the frequency for DMSO unique peaks. As expected, the common peak subset showed similar accessibility in both DMSO and MB-3 treated cells, with an average accessibility of 21.75% in DMSO and 22.83% in case of MB-3 (Fig 6.9A).

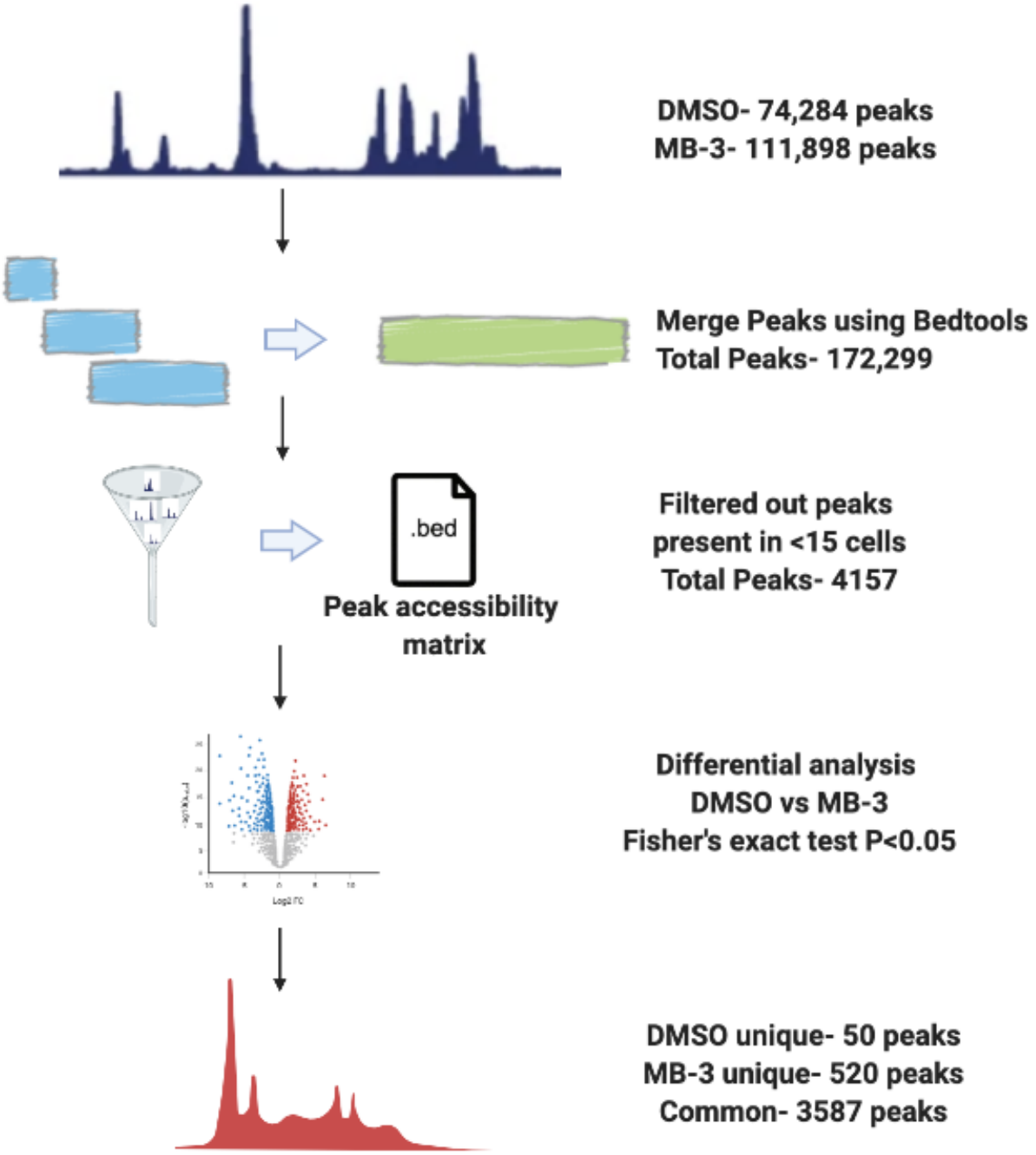


Figure 6.8: Single cell ATAC sequencing pre-processing and filtering.

Schematic for single cell ATAC sequencing analysis. Pre-processing involved merging peaks using Bedtools, filtering out peaks which may introduce technical noise and performing differential accessibility analysis to further define peaks unique to DMSO (control), unique to MB-3 (treated), and common to both population of cells.

To further characterize and understand the biological relevance of the individual subset of peaks, I annotated the peaks to nearby genes using Genomic Regions Enrichment of Annotations Tool (GREAT) (Mclean *et al.*, 2010). In control cells treated with DMSO, 41.66% of the unique peaks were found to be proximal to gene transcription start sites (TSS) (Fig 6.8B). Similarly, in MB-3 treated cells, 37.68% of peaks were proximal to TSS. However, the peaks distal to TSS (beyond 5kb), and likely may be attributed as cis-regulatory regions, displayed a higher accessibility in MB-3 treated cells where 38.69% of the distal peaks were annotated compared to 30.55% of the peaks in case of DMSO (Fig 6.9C). Overall, this also indicated that in case of control DMSO cells, there is a reduction in accessibility in distal peaks compared to proximal peaks whereas in case of MB-3 treated cells, both proximal and distal regulatory elements seem to be actively involved. This observation may indicate epigenomic reprogramming of KAT2A associated target genes. The common subset of peaks had a similar genomic region-gene association relationship as that of control cells (Fig 6.9D).

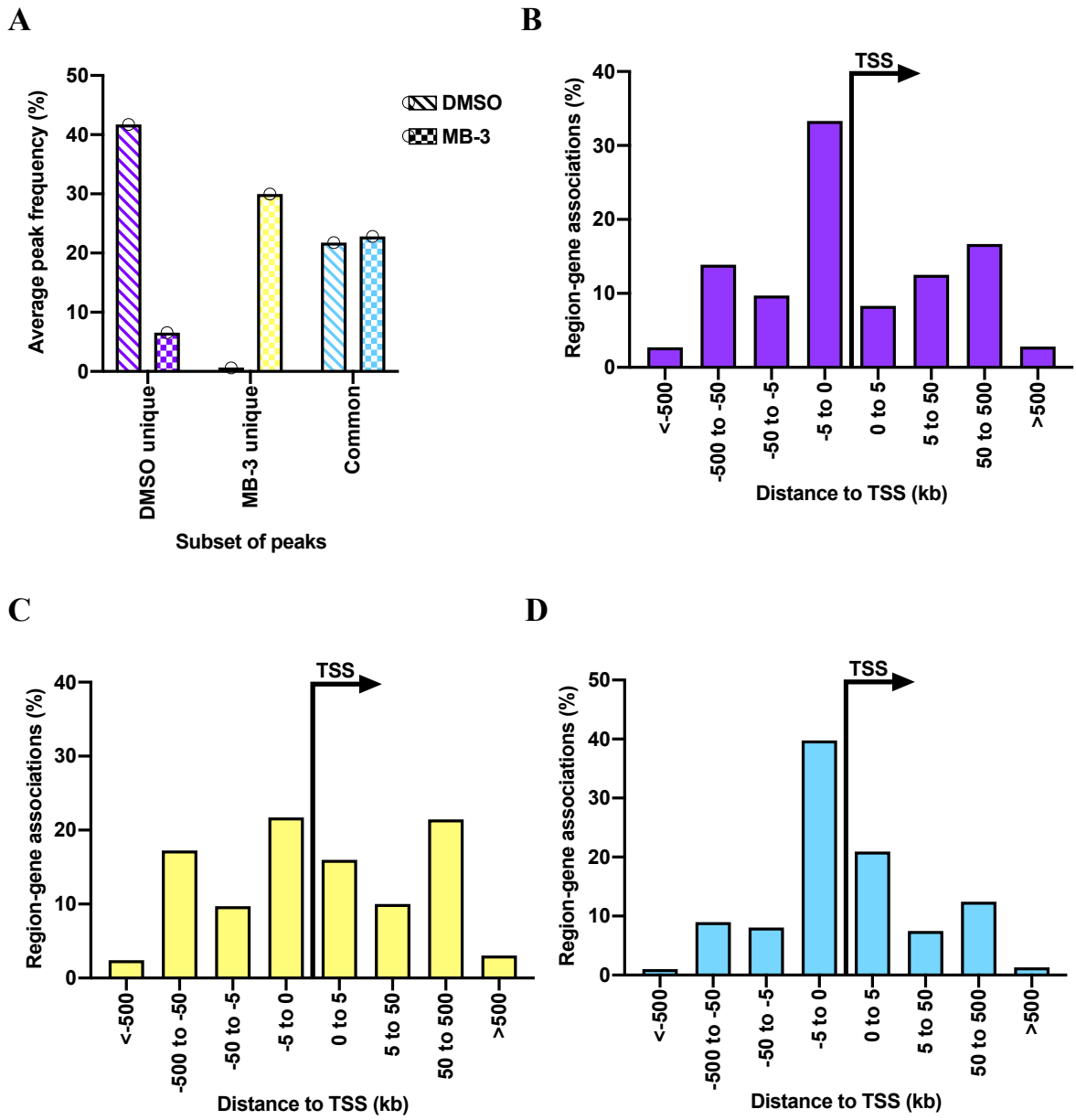


Figure 6.9: GREAT analysis for region-gene associations.

(A) Average peak frequency for DMSO unique peaks (purple), MB-3 unique peaks (yellow) and common peaks (blue) in MB-3 compared to DMSO, **(B)** GREAT analysis for DMSO unique peaks, where each peak is associated with a region defined by proximity to TSS, **(C)** GREAT analysis for MB-3 unique peaks, where each peak is associated with a region defined by proximity to TSS, **(D)** GREAT analysis for common peaks, where each peak is associated with a region defined by proximity to TSS.

6.8 MB-3 treated cells possess differential chromatin accessibility pattern

Since inhibition of KAT2A did not show any effect on global chromatin accessibility patterns in the common subset of peaks, but indicated a differential regulatory reprogramming, I visualized the cells using dimensionality reduction analysis. The combined peak accessibility matrix is binary in nature, where 1 represents an accessible region and 0 represents an inaccessible region. Therefore, to compute cell to cell distances, I followed the approach of calculating Jaccard distance (Jaccard, 1901) (Methods). Jaccard distance is a dissimilarity measure which is the ratio between the number of peaks that are unique to a cell and the total number of peaks that are open in two cells. This measure allows us to derive a pairwise distance between two cells based on similarity of peak accessibility. After calculating the dissimilarity ratio for individual pairs of cells based on the accessible peaks, the generated matrix was utilized as an input for dimensionality reduction analysis using t-distributed stochastic neighbour embedding (t-SNE) (Van Der Maaten and Hinton, 2008). The tSNE plot generated highlighted DMSO and MB-3 treated cells as part of one cluster (Fig 6.10A).

Further delving into this, I performed k-medoid clustering which employs the calculation of a central point of the cluster, termed as a medoid. The cells are further clustered based on their distance with the medoid. The medoid is calculated based on Jaccard distance as mentioned above. Upon clustering, DMSO cells grouped into a single cluster, whereas MB-3 cells formed two different clusters, indicating the presence of heterogeneity amongst the MB-3 population of cells (Fig 6.10B). The average peak frequency of MB-3 cluster I cells was ~30%, which was higher compared to control DMSO and MB-3 cluster II cells, which displayed an overall average peak frequency of ~20%. This suggested a variability in peak accessibility pattern upon inhibition of KAT2A using MB-3. This was an interesting observation and could potentially link to the differential epigenetic regulation of targets upon KAT2A inhibition.

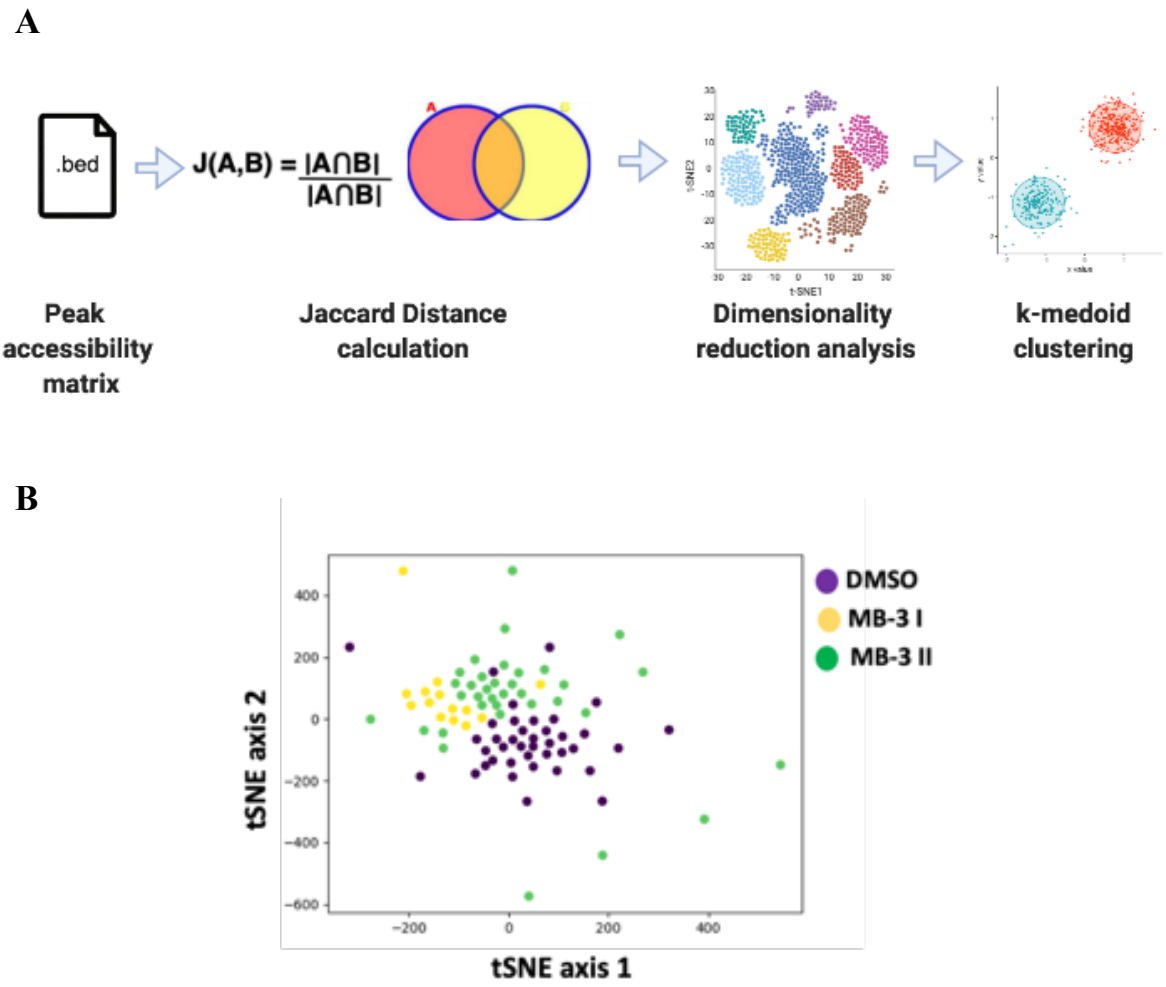


Figure 6.10: k-medoid clustering analysis.

(A) Schematic for k-medoid clustering for all cells including DMSO (control) and MB-3 (treated). The analysis started from calculating Jaccard distance, performing dimensionality reduction analysis using tSNE, and overlaying cluster membership colours obtained from k-medoid clustering onto the tSNE plot, **(B)** k-medoid clustering highlighting three different clusters, namely, DMSO (purple), MB-3 I (yellow) and MB-3 II (green).

6.9 Increase in transcriptional variability may be consequential to differential chromatin accessibility

After performing dimensionality reduction analysis and observing differential chromatin accessibility pattern with MB-3 treated cells, I wanted to study different gene ontologies which may be reprogrammed differentially upon inhibition of KAT2A. For this, I performed

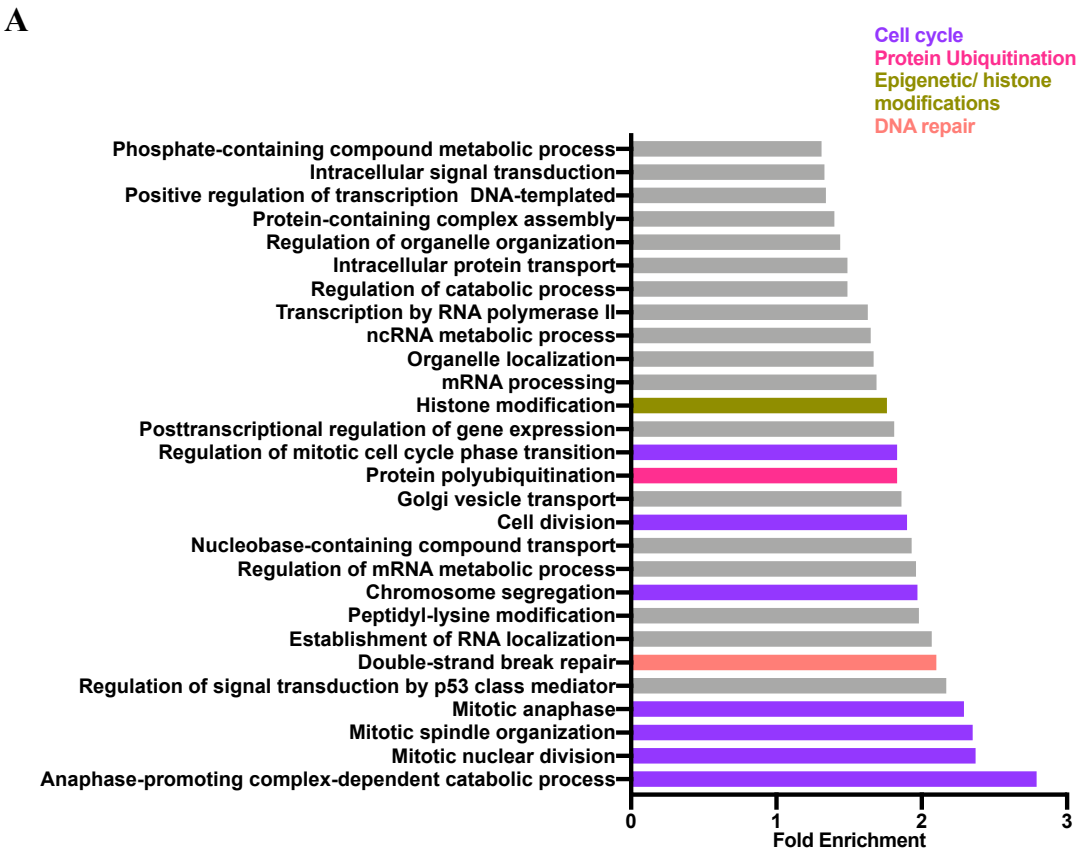
differential accessibility analysis comparing the MB-3 I cluster with DMSO and MB-3 II cluster individually, using both Fisher's exact test and information gain methods (methodology described previously). The differential peak analysis identified 3,067 peaks which were more accessible in MB-3 I cells compared to DMSO cells and 1,087 peaks more accessible in MB-3 I cells compared to MB-3 II cells.

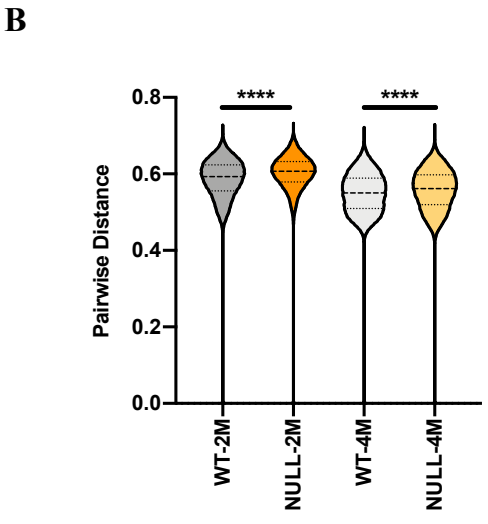
To further understand the biological meaning of the higher accessibility peaks, I annotated the peaks to their closest genes using the annotatePeaks tool in HOMER (Duttke *et al.*, 2019) (Methods). After annotating the peaks to their respective closest genes, I performed Panther gene ontology analysis for peaks that displayed higher accessibility in MB-3 cluster I compared to DMSO. This analysis revealed an enrichment in genes associated with cell cycle, DNA repair, protein ubiquitination and histone modifications (Fig 6.11A). This was an interesting observation which was compatible with the increased accessibility observed at distal enhancer regions upon inhibition of KAT2A.

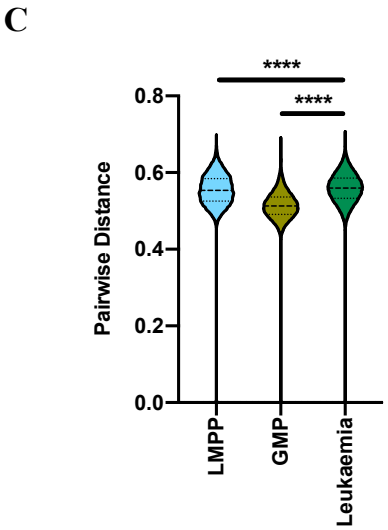
Further, I wanted to understand whether the observed epigenetic variability upon MB-3 treatment is associated with transcriptional changes. For this, I made use of scRNA-seq analysis conducted using *RUNX1-RUNX1T1(9a)* primary mouse pre-leukaemia samples. The genes associated with higher chromatin accessibility in MB-3 I cluster compared to DMSO, were converted to respective mouse gene IDs using BioMart (Kinsella *et al.*, 2011). Transcriptional variability calculations were made using pairwise correlations on this set of genes (as described previously). Global analysis suggested an increase in transcriptional variability in *Kat2a* NULL cells compared to *Kat2a* WT at respective time points with a prominent increase at 2 months (Fig 6.11B). The leukaemia progenitors compartment also showed an increase in pairwise correlation coefficient compared to the rest of the cells (Fig 6.11C), consistent with the observation in primary *RUNX1-RUNX1T1(9a)* pre-leukaemia.

Having compared MB-3 I cluster with DMSO, I looked at the differential peaks in MB-3 I compared to MB-3 II. Again, I conducted peak annotation to nearest genes using HOMER and performed Panther gene ontology analysis. The gene expression programmes which were found to be enriched were similar to the above comparison with DMSO. These were namely cell cycle, DNA repair, protein ubiquitination, and histone modifications (Fig 6.11D).

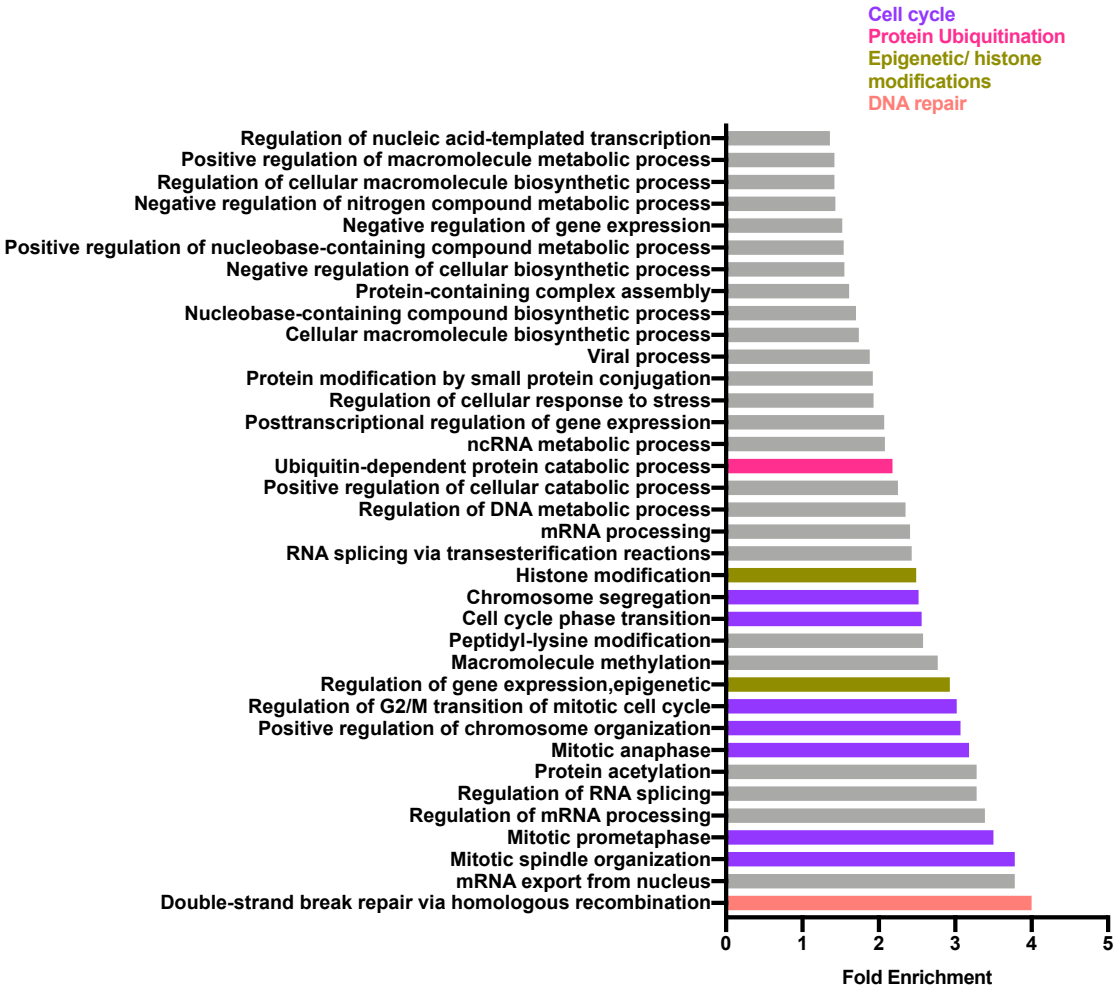
Integration with scRNA-seq data to calculate transcriptional variability indicated a similar trend of increase in pairwise distance upon *Kat2a* loss (Fig 6.11E). The leukaemia progenitor cells also showed an increase in transcriptional variability compared to the remaining population of cells (Fig 6.11F). These observations were compatible with the findings of epigenetic reprogramming upon KAT2A inhibition, and further suggest that transcriptional variability upon *Kat2a* loss may be consequential to inherent epigenetic variability.


A
Cell cycle
Protein Ubiquitination
Epigenetic/ histone modifications
DNA repair
Phosphate-containing compound metabolic process
Intracellular signal transduction
Positive regulation of transcription DNA-templated
Protein-containing complex assembly
Regulation of organelle organization
Intracellular protein transport
Regulation of catabolic process
Transcription by RNA polymerase II
ncRNA metabolic process
Organelle localization
mRNA processing
Histone modification
Posttranscriptional regulation of gene expression
Regulation of mitotic cell cycle phase transition
Protein polyubiquitination
Golgi vesicle transport
Cell division
Nucleobase-containing compound transport
Regulation of mRNA metabolic process
Chromosome segregation
Peptidyl-lysine modification
Establishment of RNA localization
Double-strand break repair
Regulation of signal transduction by p53 class mediator
Mitotic anaphase
Mitotic spindle organization
Mitotic nuclear division
Anaphase-promoting complex-dependent catabolic process
0
1
2
3
Fold Enrichment

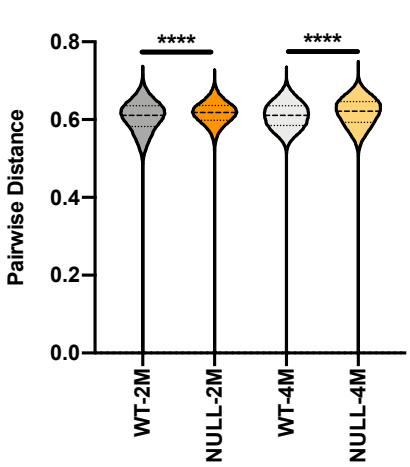

B
Pairwise Distance
0.0
0.2
0.4
0.6
0.8
****
****
WT-2M
NULL-2M
WT-4M
NULL-4M


C
Pairwise Distance
0.0
0.2
0.4
0.6
0.8
****
****
LMPP
GMP
Leukaemia

D


Cell cycle
Protein Ubiquitination
Epigenetic/ histone modifications
DNA repair
Regulation of nucleic acid-templated transcription
Positive regulation of macromolecule metabolic process
Regulation of cellular macromolecule biosynthetic process
Negative regulation of nitrogen compound metabolic process
Negative regulation of gene expression
Positive regulation of nucleobase-containing compound metabolic process
Negative regulation of cellular biosynthetic process
Protein-containing complex assembly
Nucleobase-containing compound biosynthetic process
Cellular macromolecule biosynthetic process
Viral process
Protein modification by small protein conjugation
Regulation of cellular response to stress
Posttranscriptional regulation of gene expression
ncRNA metabolic process
Ubiquitin-dependent protein catabolic process
Positive regulation of cellular catabolic process
Regulation of DNA metabolic process
mRNA processing
RNA splicing via transesterification reactions
Histone modification
Chromosome segregation
Cell cycle phase transition
Peptidyl-lysine modification
Macromolecule methylation
Regulation of gene expression,epigenetic
Regulation of G2/M transition of mitotic cell cycle
Positive regulation of chromosome organization
Mitotic anaphase
Protein acetylation
Regulation of RNA splicing
Regulation of mRNA processing
Mitotic prometaphase
Mitotic spindle organization
mRNA export from nucleus
Double-strand break repair via homologous recombination
0
1
2
3
4
5
Fold Enrichment

E


****
****
0.8
0.6
0.4
0.2
0.0
Pairwise Distance
WT-2M
NULL-2M
WT-4M
NULL-4M

F

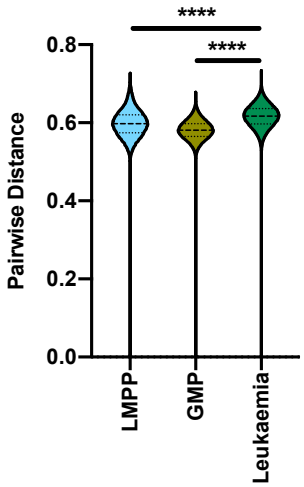

****
****
0.8
0.6
0.4
0.2
0.0
Pairwise Distance
LMPP
GMP
Leukaemia

Figure 6.11: Gene ontology analysis and integration with scRNA-seq data.
(A) Panther gene ontology analysis for genes upregulated in MB-3 I cluster compared to DMSO (p-adj<0.05*). Enriched categories include Cell cycle (purple), Protein ubiquitination (dark pink), Epigenetic/histone modifications (green) and DNA repair (light pink). **(B)** Transcriptional variability measurement for the upregulated genes in MB-3 I cluster based on pairwise correlation method using primary samples from scRNA-seq data (*Kat2a* WT 2 months as WT-2M, *Kat2a* NULL 2 months as NULL-2M, *Kat2a* WT 4 months as WT-4M, *Kat2a* NULL 4 months as WT-4M), **(C)** Transcriptional variability measurement for the upregulated genes in MB-3 I cluster based on pairwise correlation method in different population of cells identified using scRNA-seq data; LMPP, GMP and Leukaemia, **(D)** Panther gene ontology analysis for genes upregulated in MB-3 I cluster compared to MB-3 II (p-adj<0.05*). Enriched categories include Cell cycle (purple), Protein ubiquitination (dark pink), Epigenetic/histone modifications (green) and DNA repair (light pink), **(E)** Transcriptional variability measurement for the upregulated genes in MB-3 I compared to MB-3 II cluster based on pairwise correlation method using primary samples from scRNA-seq data (*Kat2a* WT 2 months as WT-2M, *Kat2a* NULL 2 months as NULL-2M, *Kat2a* WT 4 months as WT-4M, *Kat2a* NULL 4 months as WT-4M), **(F)** Transcriptional variability measurement for the upregulated genes in MB-3 I compared to MB-3 II cluster based on pairwise correlation method in different population of cells identified using scRNA-seq data; LMPP, GMP and Leukaemia.

In this chapter, I studied the stepwise progression of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT and *Kat2a* NULL pre-leukaemia cells using single cell pseudotime trajectory approaches. The analysis suggested *Kat2a* WT cells follow a linear trajectory whereas *Kat2a* NULL cells follow a branched trajectory. The multiple routes followed by *Kat2a* NULL cells are indicative of alterations in transcriptional landscape which further promote cell fate diversification during early stages of *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation. There was an observation of abrupt transformation in *Kat2a* WT cells, which could be due to inability in capturing the transformed cells at intermediate stages of transformation or could be an artifact of the algorithm implemented in the Monocle 3 tool utilized for trajectory analysis. Monocle 3 divides cells into larger and more well separated groups called partitions, using a statistical test from Alex Wolf and colleagues (Wolf *et al.*, 2019), introduced as part of their Partition-based graph abstraction (PAGA) algorithm. While learning the single cell pseudotime trajectory each of these partitions eventually forms a separate trajectory. In case of genotype specific trajectories, all of the cells had fallen within the same partition as per the algorithm, hence, a

single trajectory was obtained. However, in global trajectory analysis, by making use of default parameters, as done for genotype-specific trajectories, 3 different partitions were obtained which only allowed the algorithm to fit the trajectory within the compartment that followed a haematopoietic hierarchy, while excluding the rest of the compartments from the trajectory. To overcome this challenge, other methods for trajectory building, for example, TSCAN (Ji and Ji, 2019), could be used in future. However, it is worth considering that individual trajectory tools are based on fundamentally different algorithms (for example TSCAN utilizes minimum spanning tree whereas Monocle works on reverse graph embedding) which would make it extremely difficult to interpret and compare the analysis of one tool to the other.

Due to unavailability of published scRNA-seq datasets similar to this study, different cell populations were identified based on the average expression of cell surface makers as per the definition of these populations in previous studies (Weissman and Shizuru, 2008; Doulatov *et al.*, 2012). The presence of a higher percentage of *Kat2a* WT cells at 2 months, constituting LMPP-like population was consistent with the *in vivo* and *in vitro* observations from *RUNX1-RUNX1T1(9a)* pre-leukaemia (Chapter-3). The presence of higher percentage of 4 months samples, both *Kat2a* WT and *Kat2a* NULL within GMP compartment is compatible with transformation hierarchy, as well as the increase in myeloid progenitor population observed in pre-leukaemia cells *in vivo.*

The observation that *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells undergo B-cell and monocytic differentiation suggested that loss of *Kat2a* facilitates accessibility towards transformation prone cells by promoting cellular diversity. This was compatible with our previous observations, where *MLL-AF9* transformed *Kat2a* NULL cells had also undergone monocytic differentiation, thereby promoting accessibility towards leukaemia stem-like cells (Domingues *et al.*, 2020). Although scRNA-seq indicated the presence of B-cell differentiation at the 2 months time point which was compatible with the flow cytometry analysis, where B220 expression was found to be prominently high in *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells at plating 2 and no differences were observed between *Kat2a* WT and *Kat2a* NULL cells at plating 3. A similar observation was noted in the flow cytometry analysis for F4/80, a marker of monocytic differentiation. However, the insights from scRNA-seq suggested monocytic differentiation is present at both time points. The reason that no changes

in monocytic differentiation could be noticed at plating 3 could highlight the fact that the conditioned methylcellulose medium in which CFC cultures were maintained supports the growth of myeloid progenitors and the terminally differentiated cells may not have been maintained well in this culture. Another reason for this could be the heterogeneity in the 4 months sample where the cells captured using scRNA-seq within the monocyte compartment had the extent of transformation comparable to the cells collected at 2 months post transplantation.

The observation of accumulation of leukaemia progenitors having RUNX1-RUNX1T1 signature was in conjunction with the perpetuation of *RUNX1-RUNX1T1(9a)* transformed pre-leukaemia clones observed *in vivo.* The gradual accumulation of leukaemia progenitors with time coincided with the progressive *RUNX1-RUNX1T1(9a)* transformation. This was in-line with the flow cytometry analysis of *RUNX1-RUNX1T1(9a)* transformed pre-leukaemic *Kat2a* WT and *Kat2a* NULL cells studied at 2 months and 4 months where an increase in GFP^+c-$Kit^+Fc\gamma R^+$ population, indicative of myeloid progenitors, was observed upon *Kat2a* loss. It would have been interesting to extend the analysis by isolating these cells using FACS sorting and further transplanting into C57/BL6 mice to study the leukaemia development with and without the presence of *Kat2a*, however, due to the lack of sufficient number of cells, this analysis could not be performed.

The increase in cell-to-cell transcriptional variability was compatible with our previous observations in *MLL-AF9* model (Domingues *et al.*, 2020). However, the reduction in transcriptional variability from 2 months to 4 months could be either due to a technical artifact owing to small sample size or cellular variability contributing towards increase in transcriptional variability at the 2 months time point. Although it is difficult to dissect cell-to-cell transcriptional variability in the presence of an underlying cellular diversity, I made an attempt using the Discrete distributional differential expression (D^3E) tool (Delmans and Hemberg, 2016) which models the parameters of transcriptional bursting dynamics as an underlying cause of observed transcriptional variability. However, due to the absence of representative training simulation datasets in the D^3E model, the results could not be interpreted with confidence. Future work may involve studying these parameters of transcriptional

bursting experimentally using single molecule RNA FISH (smRNA-FISH) (Raj *et al.*, 2008) or droplet digital PCR (ddPCR) analysis (Hindson *et al.*, 2011).

The insights obtained from scATAC-seq analysis suggested that the variability in accessible chromatin landscape may further lead to an enhanced cell-to-cell transcriptional variability. However, it is worth noting that the scATAC-seq analysis was associated with few caveats. One of these was the sample size, which was smaller and perhaps quite noisy, which could be addressed in future experiments. Another caveat associated with scATAC-seq is the sparse nature of data, due to which the peaks called using single cell data are not sufficiently reliable for downstream processing. To circumvent these challenges, performing bulk ATAC-seq in parallel on the same samples would enable calling of more reliable peaks which could be utilized for downstream analysis (Schep *et al.*, 2017).

Nonetheless, the epigenomic reprogramming observed upon inhibition of KAT2A in terms of redistribution of peaks accessible upon MB-3 treatment, suggested a collaborative role of distal and proximal regions in regulating the KAT2A associated target genes. However, it is worth noting that the differentially expressed genes upon *Kat2a* loss during *RUNX1-RUNX1T1(9a)* pre-leukaemia progression, namely ribosomal biogenesis, and mitochondrial ATP biosynthetic associated genes, did not show any changes in their chromatin accessibility pattern. This could be due to the fact that these transcriptional programmes may be specifically altered during pre-leukaemia, whereas scATAC-seq data was generated from a representative established leukaemia cell line. The consistent capture of cell cycle and DNA repair machinery associated genes may suggest a cellular reprogramming dependent on differential chromatin signature or an enhanced DNA damage associated with *RUNX1-RUNX1T1(9a)* progression which are worth exploring in future as described in discussion.

7 Discussion

My work in this thesis shows the mechanistic role of cell-to-cell transcriptional variability induced upon *Kat2a* loss in governing cellular fate diversification and hence promoting *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation which can be further extended to other models of pre-leukaemia. This work represents a novel attempt at understanding the process of leukaemia initiation by introducing *Kat2a* loss as a regulator of transcriptional instability. This allowed in modelling pre-leukaemia stages and further associating the components of epigenetic variability in the form of chromatin accessibility in aiding disease progression. These observations have opened up a potential research area where the general contribution of *Kat2a* loss and consequent increase in transcriptional variability during pre-leukaemia can be further explored. These studies may further be extended to disease relapse, akin to disease initiation, where the presence of pre-leukaemia clones has been observed (Koeffler and Leong, 2017). This would further allow the identification of regulatory modes or modules that contribute to AML progression in the absence of mutation. These regulatory modules may potentially be targetable therapeutically or constitute an indicative prognostic signature that may be used in the clinic for risk stratification.

As stated earlier, AML is a heterogeneous disease entity with as many biological behaviours and natural histories as the mutation spectra observed in patients. It is worth noting that a proportion of leukaemias evolve through a prolonged pre-leukaemic phase. Understanding of the molecular mechanisms that may hinder or accelerate progression to overt leukaemia is central to the development of early diagnostic tools and early therapeutic intervention approaches. In this context, Kat2a depletion poses a fascinating model of perturbation of AML progression that can either prevent or accelerate leukaemia progression, depending on the stage of disease progression. Kat2a is central to transcriptional activation, particularly at the level of promoters (Domingues *et al.*, 2020) and in cross-talk with the basic transcriptional machinery, and functions as part of distinct histone-modifying complexes (Arede *et al.*, 2020). It is, in principle, a ‘druggable’ target (Domingues *et al.*, 2020), and detailed understanding of Kat2a mechanisms of action in different forms of AML will inform drug design to complement conventional therapeutic schemes in a disease-specific and personalised manner for increased efficacy and minimum deleterious side effects.

7.1 *Kat2a* loss may serve as a tool to study the process of transformation in other pre-malignancies

This thesis work sets up a paradigm in modelling the early steps of disease progression. The loss of *Kat2a* poses an excellent tool to understand the early steps of leukaemia progression which may further be extended to different pre-malignancies. Pre-leukaemia cell populations are marked by continuous acquisition and loss of specific mutations, sometimes occurring at different timepoints. This leads to simultaneous evolutionary convergence and divergence among particular clones and subclones during the course of disease. With the evolution of the disease, the composition of these cell populations changes remarkably, with selection occurring at different genetic and epigenetic levels. Moreover, these cell populations serve as a fantastic reservoir to understand early disruptions in epigenetic and transcriptional landscape.

In line with this, my work focussed on characterizing and modelling the pre-leukaemia stages of *RUNX1-RUNX1T1(9a)* model of leukaemia. *RUNX1-RUNX1T1* is characterized by a t(8;21) translocation event which leads to alteration of gene expression and haematopoietic cell proliferation but does not lead to development of leukaemia (Peterson and Zhang, 2004). This is in corroboration with the fact that *RUNX1-RUNX1T1* has a weakly oncogenic effect and requires collaborating mutations to develop leukaemia. On the other hand, the spliced version *RUNX1-RUNX1T1(9a)*, includes an extra exon, exon 9a, of the *RUNX1T1* gene and encodes a C-terminally truncated *RUNX1-RUNX1T1* protein of 575 amino acids. The expression of *RUNX1-RUNX1T1(9a)* is found to promote development of leukaemia in a mouse retroviral transduction–transplantation model (Yan *et al.*, 2006).

The pre-leukaemia clones in the *RUNX1-RUNX1T1(9a)* model were collected at 2 months and 4 months post leukaemia-initiation and were characterized as cell populations that were GFP^+ c-Kit^+FcγR^+. These cell populations were found to accumulate and expand within the progenitor cell compartment obtained from *Kat2a* NULL animals, suggesting that loss of *Kat2a* promotes accumulation of myeloid progenitors upon pre-leukaemia transformation. The definition of pre-leukaemia clones utilized in this study was compatible with the L-GMPs defined by Helin and colleagues in *RUNX1-RUNX1T1(9a)* leukaemia (Rasmussen *et al.*, 2015). This observation was accompanied by an increasing trend towards in the c-Kit^+ cell population

within the GFP^+ cell compartment in *Kat2a* NULL cells isolated from spleen (Student's t-test, p= 0.0642, data not shown), suggesting the presence of early progenitors. In addition to this, the overall distributions of different haematopoietic compartments in bone marrow cells including $Gr1^-Mac1^-$, c-Kit^+Sca1^-, c-Kit^+Sca1^+, and $Cd34^+Fc\gamma R^+$ cell populations were found to be altered, suggesting *Kat2a* associated changes in *RUNX1-RUNX1T1(9a)* pre-leukaemia hierarchy.

Further, these pre-leukaemia cell populations were found to be associated with an enhanced clonal expansion capacity as reflected by serial replating assay. *Kat2a* NULL bone marrow cells transformed with *RUNX1-RUNX1T1(9a)* displayed a marked increase in self-renewal potential suggesting these cells have the capability to maintain the transformants which may have further expanded and consequently led to accelerated leukaemia progression. This was compatible with an *in vitro* colony forming assay performed on *RUNX1-RUNX1T1(9a)*$^+$ transformed *Kat2a* WT and *Kat2a* NULL cells, which also showed an increase in replating capacity in *Kat2a* NULL cells. Interestingly, in a maintenance set-up where *Kat2a* knockout was performed post *RUNX1-RUNX1T1(9a)* transformation, self-renewal potential was significantly reduced. This observation was in line with a previous study conducted in our lab (Domingues *et al.*, 2020) where reduction in colony-forming potential was observed in *MLL-AF9* induced leukaemia upon *Kat2a* loss, suggesting contrasting roles of *Kat2a*. These observations highlight that *Kat2a* impacts self-renewal in a stage-specific manner where its loss in transformed leukaemic cells decreases colony forming potential, whereas its loss in a disease initiation context enhances replating capacity.

These pre-leukaemia cells were further found to have a linear increase in White Blood Cell (WBC) count consistently in both *Kat2a* WT and *Kat2a* NULL genotypes, which was compatible with the first phenotypic characterization study of *RUNX1-RUNX1T1(9a)* leukaemia (Yan *et al.*, 2006). No changes were observed in haemoglobin level and platelet counts during the course of leukaemia development except a sharp decrease in *Kat2a* NULL animals at 44 weeks post transplantation. This may be due to aging or could be technical and may not be a consequence of pre-leukaemia burden as it does not show any association with changes in WBC levels.

The *RUNX1-RUNX1T1(9a)* leukaemias which developed were found to be in accordance with the classification of the haematopathology subcommittee of the Mouse Model of Human Cancers Consortium (MMHCC) (Kogan *et al.*, 2002). The clinical characteristics of these mice were suggestive of AML without maturation. In line with the literature, the leukaemia cells obtained from *RUNX1-RUNX1T1(9a)* did not show a typical distribution of the three myeloid progenitor populations, including common myeloid progenitors (FcγRIII^{lo}CD34^{+}), granulocyte and macrophage progenitors (FcγRIII^{hi}CD34^{+}), and megakaryocyte and erythroid progenitors (FcγRIII^{lo}CD34$^{–}$) within the progenitor compartment (c-Kit^{+} Sca-1$^{–}$) (Yan *et al.*, 2006) (Yan *et al.*, 2004). All developed leukaemias had an unique profile highlighting different subset of markers confirming the multi-hit hypothesis of *RUNX1-RUNX1T1(9a)* leukaemia development (TCGA, 2013b).

Our retroviral transduction-based transplantation model of *RUNX1-RUNX1T1(9a)* showed the first *Kat2a* WT animal to have developed leukaemia within 171 days post transplantation. This was compatible with available literature suggesting the development of *RUNX1-RUNX1T1(9a)* leukaemia within a timeframe of 120-200 days post transplantation (Bäumer *et al.*, 2014)(Hatlen *et al.*, 2016)(Kobayashi *et al.*, 2017). However, lower transduction efficiency in my transplantation experiments may have contributed towards longer latency period of *RUNX1-RUNX1T1(9a)*-induced leukaemogenesis. Upon loss of *Kat2a*, an acceleration in leukaemia progression was observed, with the first animal showing signs of leukaemia after 104 days of transplantation. It is worth noting that my experiment was conducted in two separate cohorts of animals with separate transduction experiments conducted at different time frames. Although the transduction efficiency was variable between the two experimental set ups, nonetheless, in both cases a consistent acceleration in leukaemia progression in *Kat2a* NULL animals transformed with *RUNX1-RUNX1T1(9a)* overexpression plasmid was observed. Since the animals were only analysed until 1-year post-transplantation and the experiment was terminated thereafter, some *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT animals which may have started developing signs of leukaemia could not be analysed.

RUNX1-RUNX1T1(9a) transformed leukaemic cells highlighted the presence of 20-30% of c-Kit^{+}Sca1^{-} population within GFP^{+} cells in bone marrow, indicative of *RUNX1-RUNX1T1(9a)* transformed early progenitors, compatible with the observation from pre-leukaemic animals.

This was in line with the observation in literature where the proportion of myeloid progenitors ($Lin^{-}Sca\text{-}1^{-}c\text{-}Kit^{+}$) was greatly increased in GFP^{+} cells. Further, most GFP^{+} cells were found to have no expression of lineage markers associated with myeloid lineage like Gr-1 and Mac1, compatible with previous studies (Yan *et al.*, 2006; Zuber *et al.*, 2009).

The *RUNX1-RUNX1T1(9a)* leukaemia burden studies for both *Kat2a* WT and *Kat2a* NULL animals showed that the transplanted mice had developed signs of anaemia, high white blood cell (WBC) count, abnormal differential counts and signs of cachexia and laboured breathing at the time of culling. These mice which developed leukaemia showed pale femurs, enlarged spleens and in some cases enlarged livers, but there was no indication of any clinical abnormalities in thymuses or lymph nodes. There was a high number of blast cells which was observed in the peripheral blood at terminal time point both in bone marrow and spleen (data not shown) compatible with previous characterization of *RUNX1-RUNX1T1(9a)* induced leukaemia burden (Yan *et al.*, 2006) (Yan *et al.*, 2004). No significant differences were observed in terms of overall leukaemia burden in *Kat2a* NULL relative to *Kat2a* WT animals, compatible with previous lab findings in case of *MLL-AF9* leukaemia, where both *Kat2a* WT and *Kat2a* knockout (KO) leukaemic animals had similar disease burdens at the point of culling (Domingues *et al.*, 2020). However, there was increasing trend in spleen size in *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL animals, perhaps indicating an early disease infiltration event which was merely strengthening the notion of accelerated *RUNX1-RUNX1T1(9a)* leukaemia development upon loss of *Kat2a.*

Overall, the combination of pre-leukaemia and leukaemia analysis of *RUNX1-RUNX1T1(9a)* in a *Kat2a* conditional knockout background suggested a perpetuation of pre-leukaemia clones upon *Kat2a* loss. These pre-leukaemia clones, as defined by $GFP^{+}c\text{-}Kit^{+}Fc\gamma R^{+}$, showed an enhanced self-renewal potential and aided in accelerated leukaemia development in absence of *Kat2a*. Going forward, it would be interesting to explore the role of *Kat2a* in other pre-leukaemia models specifically the ones initiated by RUNX1 mutation.

One such model which could benefit from *Kat2a* based perturbation studies is Myelodysplastic syndrome (MDS). Myelodysplastic syndromes (MDS) are clonal disorders of haematopoietic stem cells characterized by ineffective haematopoiesis. The disorder is associated with

abnormal myeloid differentiation leading to chronic cytopenias (reduction in number of mature blood cells) and to a variable risk of progression to AML (Itzykson and Fenaux, 2013). Even though the presence of various chromosomal abnormalities, mutations, and epigenetic changes have been observed in MDS progenitors, the earliest cellular stages at which pathogenic events occur is still not clear. Some studies in MDS have focused on the subset of patients with chromosomal 5q deletion (5q−) and have shown that stem cells in MDS harbour the deletion (Nilsson *et al.*, 2007). Another study showed that these cells persist in the bone marrow (BM) of patients with 5q− during the process of treatment and can be predictive of relapses (Tehranchi *et al.*, 2010). The number of recurrently mutated genes is greater in more advanced MDS subtypes, and acquisition of certain mutations including RUNX1 is associated with transformation to AML (Bejar *et al.*, 2011). Splicing machinery associated genes have been found to be one of the most frequently impaired genes in MDS which can be identified in >50% of MDS patients (Bejar and Steensma, 2014). This is in line with my observations from individual analysis performed in both *Kat2a* WT and *Kat2a* NULL cells transformed with *RUNX1-RUNX1T1(9a),* where a downregulation of genes associated with splicing machinery was found to be associated with AML transformation. In addition to this, aberrant expression and mutation of multiple ribosomal proteins has been observed in MDS (Pellagatti *et al.*, 2008) (McGowan *et al.*, 2011) where rRNA expression is decreased in haematopoietic progenitor cells from MDS patients (Raval *et al.*, 2012), compatible with my observations in *RUNX1-RUNXT1(9a)* model upon *Kat2a* loss, suggesting that Kat2a loss can serve as a tool in modelling the early stages of 5q mediated MDS transformation.

Another pre-leukaemic condition which has a predisposition towards different haematological and epithelial malignancies is Fanconi anaemia (FA). FA is a rare genetic condition characterized by congenital abnormalities, chromosome fragility, progressive bone marrow failure, and cancer susceptibility (Quentin *et al.*, 2011). Most FA patients develop bone marrow failure throughout the course of the disease, usually during their first and second decades of life, and for the majority of patients the suspicion of FA is only made after the onset of pancytopenia (Shimamura and Alter, 2010). A strong predisposition towards haematological malignancies has been observed with cumulative probabilities of an FA patient developing MDS or AML being 30% to 40% by the age of 40 years (Alter, 2003). In a number of patients, the underlying diagnosis of FA is not known until an MDS/AML occurs, highlighting the

urgent need of biomarker identification for disease prognosis. Epigenetic studies on FA have been mostly conducted on the role of histone deacetylation as well as hypomethylation in affecting FA-BRCA pathway during disease progression suggesting a dependency on epigenetic modifiers for disease progression (Belo *et al.*, 2015). These observations support the notion that Kat2a loss may serve as a tool to study this pre-malignancy.

Another model representing pre-leukaemia is ETV6-RUNX1, also known as TEL-AML1, one of the most frequent chromosomal translocations involving t(12;21)(p13;q22) in childhood Acute Lymphoblastic Leukaemia (ALL) with an incidence of approximately 25% (Sun, Chang and Zhu, 2017). This translocation occurs in haematopoietic stem cells and leads to the establishment of pre-leukaemic progenitor B-cell clones that persist in the bone marrow for several years. It has been shown that the translocation itself is insufficient to generate ALL and that secondary somatic mutations are necessary to activate the leukaemic phenotype (Sun, Chang and Zhu, 2017). The detection of the ETV6-RUNX1 fusion sequence in identical twins and in neonatal blood spots of children with ALL indicate that this gene fusion originates in the prenatal period (Ford *et al.*, 1998). The transcription factor binding properties of ETV6-RUNX1 have been mostly explored in context of RUNX1 where the binding of ETV6-RUNX1 to RUNX1 target genes has been shown to potentially convert RUNX1 to a negative transcriptional regulator (Zelent, Greaves and Enver, 2004). A recent study by Enver group highlighted the gene expression programmes associated with ETV6-RUNX1 maintenance (cell cycle, E2F targets, MYC targets) (Wray *et al.*, 2020) which coincided with the transcriptional programmes impaired upon *Kat2a* loss, suggesting that Kat2a loss may serve as an excellent tool to study this pre-leukaemia model.

In parallel to this, in this thesis, pre-leukaemia studies were also conducted using the *Idh1*R132H model developed by our collaborator Prof. George Vassiliou's laboratory at WT-Sanger Institute, which we crossed into our conditional *Kat2a* knockout background. *IDH1* mutations have been identified as recurrent mutations during leukaemogenesis (Mardis *et al.*, 2009). The presence of these mutations is mutually exclusively of the transcription-factor fusions, suggests that they may have functional implications in the initiation of AML (TCGA, 2013b; Papaemmanuil *et al.*, 2016). Studies conducted on *Idh1*R132H are quite sparse, and the only mouse model published till date did not report progression to acute leukaemia (Sasaki *et*

al., 2012b), indicating that additional collaborating mutations are required. Our analysis of *Idh1*R132H pre-leukaemia was also compatible with the literature, where an enhanced WBC count was observed but progression to acute leukaemia was not observed. However, the transplanted animals showed an accumulation of early progenitors upon *Kat2a* loss, characterized by c-Kit^{+}Mac1^{-}, compatible with the perpetuation of pre-leukaemic clones observed in *RUNX1-RUNX1T1(9a)* pre-leukaemia.

The pre-leukaemia cells transformed with *Idh1*R132H were analysed post 4 weeks and 20 weeks of leukaemia initiation. Upon phenotypic characterization of pre-leukaemia clones, no significant differences between the two genotypes were found in any of the haematopoietic compartments including HSCs, suggesting that loss of *Kat2a* at early stages of pre-leukaemia progression does not impact the BM cellularity. The KSL and KL population of cells followed a similar trend where no changes upon *Kat2a* loss were seen, however, an increasing trend in KL population was observed with time, in line with the previous literature suggesting accumulation of early progenitors during pre-leukaemia progression (Sasaki *et al.*, 2012b). Further, these pre-leukaemia cell populations at 4 weeks were found to be associated with an enhanced clonal expansion capacity as reflected by serial replating assay, where *Kat2a* NULL bone marrow cells transformed with *Idh1*R132H had a marked increase in self-renewal potential suggesting these cells have the capability to maintain the transformants. Strikingly, no such differences were observed at 20 weeks post leukaemia initiation, suggesting that the enhanced colony forming capability of *Kat2a* NULL cells is an early effect upon *Kat2a* loss. Overall, *Kat2a* depletion serves as an excellent model to study early stages of pre-leukaemia progression as seen in case of *RUNX1-RUNX1T1(9a)* and *Idh1*R132H models which can be further extended to other pre-malignancy models including disease relapse.

It is worth noting that the definition of pre-leukaemia is not strictly restricted to the cell populations having increased incidence of evolution into AML but also corresponds to the pre-leukaemic clones which can exist at the time of complete morphologic remission after standard aggressive chemotherapy for AML (Koeffler and Leong, 2017). This phenomenon was first observed by Phil Fialkow and Claus Bartram in the early 1990s where they noted that about 25% of the AML patients who achieved complete morphologic remission by chemotherapy continued to have an abnormal clone as measured by X-inactivation (Fialkow, Janssen and

Bartram, 1991). Since then, several high throughput sequencing studies have shown that individuals whose AML cells at diagnosis had mutations of either DNMT3A, IDH2, TET2, UAF1, or SRSF2 often had the same alteration at complete morphologic remission (Corces-Zimmerman *et al.*, 2014c; Garg *et al.*, 2015; Klco *et al.*, 2015; Sun *et al.*, 2017). Surprisingly, the driver mutations in genes such as FLT3, RUNX1, RAS, NPM1, CEBPA, and WT1, although present at diagnosis of AML, were no longer found in complete remission. Ley and colleagues had reported that ~50% of the individuals whose AML cells carried DNMT3A, TET2, or IDH1/2 mutations at diagnosis continued to have the same mutation in >5% of the haematopoietic cells at complete morphologic remission, but the classical driver mutations (for example, NPM1, FLT3 and RAS) found at diagnosis were not detectable at remission (Klco *et al.*, 2015). Another study performed by Dick's group on a cohort of individuals whose AML blast cells had mutations of DNMT3A and NPM1 at both diagnosis and relapse indicated the presence of only the DNMT3A mutation upon morphologic complete remission (Shlush *et al.*, 2014b). In the same study, human bone marrow cells at diagnosis of AML were injected into immunodeficient mice where a total of 37% of samples did not engraft, 40% of samples recapitulated the AML blasts in the mice, but most interestingly, 23% of the samples produced multi-lineage engraftment that appeared normal. These multilineage engrafts had continued to have mutations of either DNMT3A or IDH1/2 suggesting that a preleukaemic clone exists even at diagnosis of AML(Shlush *et al.*, 2014b). These observations highlight that Kat2a loss may serve as a tool in understanding the molecular mechanisms behind disease relapse models.

7.2 ***Kat2a*** **loss may serve as a tool to study mitochondrial metabolism**

The global differential gene expression analysis in combination with comparisons performed at individual time points indicated an enrichment in translation/ribosome biogenesis and mitochondrial ATP synthesis pathways upon *Kat2a* loss. These observations were compatible with a previous study conducted in our lab on *MLL-AF9* leukaemia where loss of *Kat2a* downregulated gene expression pathways associated with translation/ribosome biogenesis and mitochondrial ATP synthesis (Domingues *et al.*, 2020). These findings altogether suggested that loss of *Kat2a* may exert its effects through general pathways rather than leukaemia specific ones strengthening the notion that *Kat2a* perturbation studies may not be restricted to pre-

malignancies and can be further extended to study the impairment in ribosomal biogenesis and mitochondrial energetics in different malignancies.

The time series comparison performed on *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells suggested the presence of transcriptional instability in gene expression programmes associated with alternative splicing and cis-splicing, along with cell cycle regulation, protein modifications, and chromatin organization. The overall analysis highlighted that loss of *Kat2a* impacts translation/ribosomal biogenesis and mitochondrial ATP synthesis during early stages of pre-leukaemia transformation and potentially need those during pre-leukaemia maintenance. However, once the *RUNX1-RUNX1T1(9a)* pre-leukaemia is established, disease progression is dependent on alterations in cell cycle progression, post-transcriptional modifications, and epigenetic instability in the form of chromatin reorganization. These insights were in agreement with literature where corruption of cell cycle regulation by *RUNX1-RUNX1T1* is found to mediate leukaemia progression (Martinez-Soria *et al.*, 2018). Previous work from the Bonifer group has also indicated the regulation of cell cycle genes and epigenetic reprogramming in terms of histone acetylation and transcription factor binding by RUNX1-RUNX1T1 in t(8;21) leukaemia (Ptasinska *et al.*, 2012). A recent preprint from Heidenreich lab in collaboration with Bonifer lab also suggested that RUNX1-RUNX1T1 affects alternative splicing of RNA in the t(8;21)-positive Kasumi-1 AML cells by controlling the choice of transcriptional start sites and by modulating expression of splicing components, suggesting a role for alternative splicing in RUNX1-RUNX1T1 leukaemogenesis (Grinev *et al.*, 2021).

Contrastingly, *Kat2a* WT cells were found to be mostly dependent on mitochondrial bioenergetics for pre-leukaemia transformation, highlighting the general significance of metabolic reconfiguration during pre-leukaemia transformation. From mitochondrial activity analysis of both *Kat2a* WT and *Kat2a* NULL cells undergoing transformation with *RUNX1-RUNX1T1(9a),* an overall reduction in active mitochondrial mass was observed in *Kat2a* NULL cells relative to *Kat2a* WT cells. This reduction was not accompanied by any changes in mitochondrial potential. Further, the proportion of cells with low mitochondrial content was found to be higher within *Kat2a* NULL cells relative to *Kat2a* WT cells undergoing transformation with *RUNX1-RUNX1T1(9a).* This was in line with the inferences from differential expression analysis of scRNA-seq data obtained from *RUNX1-RUNX1T1(9a)* pre-

leukaemia, indicating that reduction in mitochondrial activity upon *Kat2a* loss aids in *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation.

The observation of reduced mitochondrial content during *RUNX1-RUNX1T1(9a)* pre-leukaemia was in line with the Warburg theory laid down in 1924 (Hay, 2016a), which stated that cancer cells are highly dependent on glycolysis not so much for energy production, but rather as a platform to produce building blocks including nucleotides, amino acids, and electron carriers that are necessary for cancer progression (Hay, 2016b). The Warburg effect ensures a constant high activity of glycolysis and maintains the balance between energy supply and the anabolic functions of glycolysis and its deviations. This was in contrast with some of the findings on chronic myeloid leukaemia (CML) as well as AML where higher mitochondrial biogenesis and higher dependency on mitochondrial oxidation in leukaemia myeloblasts was observed compared with normal haematopoietic precursor cells (Green *et al.*, 2010; Vakana *et al.*, 2011; Farge *et al.*, 2017; Mattes, Vellenga and Schepers, 2019). Strikingly, a recent study presented a slight variation to this notion (Marsin *et al.*, 2000). The study suggested that decreased energy levels in cells activates AMP-activated protein kinase (AMPK), a crucial energy sensor, which promotes ATP production by increasing the activity and expression of proteins involved in catabolic processes including glycolysis and fatty acid oxidation, while conserving ATP by switching off biosynthetic pathways such as fatty acid, glycogen, and protein synthesis (Marsin *et al.*, 2000)(Marsin *et al.*, 2002). On these lines, a similar study conducted on *MLL-AF9* leukaemia indicated that dietary restriction in mice harbouring *MLL-AF9*-induced AML leads to activation of AMPK, which is essential to support leukaemia development. Thus, AMPK deletion significantly delays leukaemogenesis and depletes leukaemia-initiating cells by reducing glucose uptake and increasing oxidative stress and DNA damage. Notably, AML-initiating cells are particularly dependent on AMPK to suppress oxidative stress in the hypoglycaemic bone marrow environment and AMPK inhibition synergizes with dietary restriction to suppress leukaemogenesis (Saito *et al.*, 2015). These studies altogether highlight that reduction in metabolic activity may be a hallmark of pre-leukaemia transformation which may aid in maintenance of leukaemia-initiating cells.

In this line, the impact of *Kat2a* on mitochondrial activity was also studied in *Kat2a* WT and *Kat2a* NULL cells *in vitro* transformed with *MLL-AF9.* Again, an overall reduction in active

mitochondrial mass was observed in *Kat2a* NULL cells. This was accompanied by an increase in relative proportion of Mito lo population. Along with alterations in mitochondrial mass, there was a subtle difference in mitochondrial membrane potential, where loss of *Kat2a* led to a reduction in mitochondrial membrane potential as read by fluorescent signal obtained from Mitostatus TMRE. These findings were in line with the observations from apoptosis analysis performed in *MLL-AF9* transformed primary cell lines where an enhanced proportion of cells undergoing apoptosis were seen upon loss of *Kat2a* (Chapter-2). It is interesting how *Kat2a* mediated mitochondrial activity alterations may impact different models of leukaemia with similar effect upon *Kat2a* loss. Although these experiments both in *RUNX1-RUNX1T1(9a)* as well as *MLL-AF9* models would have strongly benefited from including another replicate, in case of both models, the increasing trend in Mito lo cell population upon *Kat2a* loss seemed quite consistent suggesting that metabolic reconfiguration observation upon *Kat2a* loss may play an important role in context of both leukaemia initiation as well as maintenance model systems.

A deeper insight into the biology of mitochondrial activity alterations was obtained by segregating the *MLL-AF9* transformed cells into two categories- one with high active mitochondrial mass (Mito hi) and the other with lower active mitochondrial mass (Mito lo). An overall reduction in number of colonies was observed in *Kat2a* NULL cells compared to *Kat2a* WT cells, compatible with the global findings discussed in Chapter-6. The cells from Mito lo fractions formed less colonies compared to the respective Mito hi fractions for both genotypes, indicating that segregation of cells based on mitochondrial content may not have a genotype-specific effect. However, the colony forming potential of *Kat2a* WT Mito lo was found to be comparable with that of *Kat2a* NULL Mito hi and *Kat2a* NULL Mito lo, overall suggesting that *Kat2a* NULL phenocopies *Kat2a* WT Mito lo.

In order to associate mitochondrial content with colony forming potential, I looked at the composition of these colonies based on a study from Lavau and colleagues (Lavau *et al.*, 1997) defining types of colonies. The analysis showed a similar trend of reduction in compact colonies, the ones without any halo of migrating cells and thus representative of immature myeloid cells, as was observed globally in *Kat2a* NULL compared to *Kat2a* WT cells, in line with the observations discussed in Chapter-6. Once again, the relative proportion of compact

colonies from *Kat2a* WT Mito lo was found to be comparable with that of *Kat2a* NULL Mito hi and *Kat2a* NULL Mito lo, overall suggesting that *Kat2a* NULL phenocopies *Kat2a* WT Mito lo, a trend compatible with total colonies. Furthermore, there was a significant reduction in mixed colony proportion, the colonies having a compact centre with a diffuse halo of differentiating progenitors, which was seen in *Kat2a* NULL Mito lo cells relative to *Kat2a* NULL Mito hi cells. The proportion of mixed colonies in *Kat2a* NULL Mito lo was comparable to the mitochondrial fractions in *Kat2a* WT cells. The proportion of dispersed colonies, comprising of large diffuse colonies without a clear centre, was the lowest in all the samples, consistent with the previous observations (Chapter-6). However, a trend towards an increase in the proportion of dispersed colonies was observed in case of *Kat2a* WT Mito lo, comparable with global *Kat2a* NULL cells, again indicating that *Kat2a* NULL phenocopies *Kat2a* WT Mito lo. Overall, the analysis allowed association of colony forming potential with the population of cells segregated on the basis of mitochondrial mass, compatible with previous observations by Simsek et al., which was based on *in vitro* colony forming assays and *in vivo* long-term repopulation assays, suggesting that separation of cells solely based on their metabolic profile markedly enriches for HSCs (Simsek *et al.*, 2010).

In line with this, the insights generated from single cell clonal cultures derived from individual mitochondrial fractions sorted from each genotype suggested a reduction in clonal expansion potential in *Kat2a* NULL cells relative to *Kat2a* WT cells transformed with *MLL-AF9*. This was compatible with the insights from colony forming assay described above as well as in Chapter-6. The cells from Mito lo fractions had reduced clonal expansion capacity compared to the respective Mito hi fractions for both genotypes, indicating that segregation of cells based on mitochondrial content may not have a genotype-specific effect. This was again in accordance with the colony forming assay observations described above, where both Mito lo fractions had reduced colony forming potential compared to their respective Mito hi fractions. This was also in line with the flow cytometry analysis (data not shown) where the viability of Mito lo fractions was compromised compared to respective Mito hi fractions suggesting that Mito lo fractions are hard to sustain in culture *in vitro.* Furthermore, the ability of a single cell to expand clonally was found to be comparable in *Kat2a* WT Mito lo with *Kat2a* NULL Mito total, again suggesting that loss of *Kat2a* may potentially phenocopy some of the aspects of *Kat2a* WT Mito lo population of cells. This was again compatible with the self-renewal assay

analysis discussed above. These analyses altogether highlighted that loss of *Kat2a* leads to reduction in active mitochondrial mass which further impacts self-renewal potential and clonogenic expansion capacity of *MLL-AF9* transformed cells *in vitro*. This was in line with the suggestion from literature that leukaemia stem cells showed increased mitochondrial mass with an increase in oxygen consumption and ATP production. These cells further exhibit increased fatty acid oxidation, consistent with high oxidative phosphorylation (OXPHOS) status and reduction in mitochondrial activity which led to enhanced anti-leukaemic effects (Bosc, Selak and Sarry, 2017). This was consistent with depletion of *MLL- AF9* leukaemia stem-like cells in our study as a consequence of reduced mitochondrial activity upon *Kat2a* loss (Domingues *et al.*, 2020).

Notably, the observed reduced mitochondrial activity in *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL pre-leukaemia cells may indicate alterations in metabolic landscape whereby the *RUNX1-RUNX1T1(9a)* transformed cells showed a reduction in mitochondrial respiration. This reduced dependency on mitochondrial respiration, in contrast to the above highlighted studies, may be associated with an enhanced cytoplasmic glycolysis (Simsek *et al.*, 2010). Previous studies have indicated high glucose uptake rate displayed by AML cells (Cunningham and Kohno, 2016). In addition to this, studies have suggested that high levels of aerobic glycolysis at diagnosis were predictive for improved therapy response and survival of a small series of AML patients (Herst *et al.*, 2011). Furthermore, a study conducted by Chen et al. showed that sera obtained from AML patients have a distinct glucose metabolic signature exhibiting significant alterations in six metabolites in this pathway. Among those, lactate, 2-oxoglutarate, pyruvate, 2-HG, and glycerol-3-phosphate were found to be negatively associated with survival. There were no significant differences among distinct WHO AML subtypes, suggesting that this metabolic signature may be a consistent feature of AML independent of cytogenetic risk groups (Chen *et al.*, 2014). Specifically, AML t(8;21) has been described to depend on glycolysis for its survival, which is further supported by the fact that the t(8;21) translocation positive Kasumi-1 cell line is highly sensitive to glycolysis inhibition, suggesting subtype specificity (Suganuma *et al.*, 2010; Isa *et al.*, 2018). A recent study further strengthened this notion by highlighting the role of ZBTB7A mutation frequently present in AML with t(8;21) translocation and correlated the loss of ZBTB7A with enhanced glycolysis rate in RUNX1-RUNX1T1 leukaemogenesis (Redondo Monte *et al.*, 2020). This study showed

that loss of ZBTB7A increases the expression of SLC glucose transporter genes as well as ENO2, PGM2, and PGM3, increasing glycolysis and sensitizing to glycolysis inhibition. Interestingly, inhibition of mitochondrial respiration revealed a profoundly increased glycolytic reserve in ZBTB7A knockout cells. They further showed that ZBTB7A can counteract RUNX1-RUNX1T1-dependent progenitor cell expansion through repression of glycolysis, opening up avenues for a targeted treatment of AML t(8;21) with metabolic inhibitors (Redondo Monte *et al.*, 2020). However, it is worth noting that the reduced mitochondrial activity of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells may also be attributed to cells being in quiescent state and indicate their presence in disease relapse (Koeffler and Leong, 2017).

Mitochondrial oxidative phosphorylation represents another aspect of mitochondrial bioenergetics which was found to be impaired upon *Kat2a* loss during *RUNX1-RUNXT1(9a)* pre-leukaemia. Mitochondrial DNA (mt-DNA) is composed of a double-stranded circular genome 16.6 kb in length without introns (Lang, Gray and Burger, 1999). It encodes 2 rRNAs, 22 t-RNAs, and 13 of the 90 proteins in the mitochondrial respiratory chain. The 13 mt-DNA-encoded proteins are translated by mitochondrial ribosomes within the mitochondrial matrix (Gaur *et al.*, 2008). Mitochondrial ribosomes differ from eukaryotic cytosolic ribosomes in their structure and chemical properties (O'Brien, 2003). In addition, they use unique protein translation machinery including distinct initiation and elongation factors. These 13 mt-DNA-encoded subunits of the electron transport chain are important for functional regulation of oxidative phosphorylation (Fukuda *et al.*, 2007).

In line with this, in this thesis, the role of mitochondrial translation was studied in both *MLL-AF9* and *RUNX1-RUNXT1(9a)* transformed *Kat2a* WT cells. For this, tigecycline, an inhibitor of mitochondrial translation was used. Tigecycline is a third generation tetracycline antibiotic that interferes with translation by blocking the interaction of aminoacyl-tRNA with the A site of the ribosome in mitochondria (as well as bacteria and eukaryotes) and has been proposed as a potential therapy in AML (Škrtić *et al.*, 2011) (Lamb *et al.*, 2015). The *MLL-AF9* transformed *Kat2a* WT cells didn't show any difference in colony forming potential upon treatment with Tigecycline, indicating that these cells may not be sensitive to mitochondrial translation inhibition. However, the colony composition of the treated cells was found to be altered, where

a significant reduction in compact colonies representative of the colonies without any halo of migrating cells was observed, highlighting the presence of immature myeloid cells. These observations suggested that *MLL-AF9* transformed cells may not be sensitive to tigecycline in terms of impacting the overall colony numbers, however, tigecycline treatment impacts the maintenance of self-renewal potential of these cells, highlighted by the reduced number of compact colonies. This reduction in compact colonies was accompanied by a gain in the proportion of non-compact colonies indicating that *MLL-AF9* transformed *Kat2a* WT cells upon inhibition of mitochondrial translation activity may mimic some of the characteristics of *Kat2a* NULL cells. Thus, the inhibition of mitochondrial translation impacts the self-renewal potential of *MLL-AF9* transformed *Kat2a* WT cells which is compatible with previous literature findings where tigecycline treatment impacts self-renewal of TEX and M9-ENL1 AML cell lines and leads to a reduced repopulating ability of these treated cells observed in NOD/SCID Nonobese diabetic/severe combined immunodeficiency (NOD/SCID) mice (Škrtić *et al.*, 2011).

The analysis was further extended to *RUNX1-RUNX1T1(9a)* leukaemia initiation set-up, where *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT cells upon treatment with tigecycline, showed reduced number of colonies at each plating suggesting that the cells undergoing transformation with *RUNX1-RUNX1T1(9a)* are sensitive to the inhibitor. This was consistent with previous studies (Rashkovan and Ferrando, 2019). Interestingly, post plating 3, which is when the percentage of *RUNX1-RUNX1T1(9a)* transformed cells is >99%, the number of colonies upon tigecycline treatment remained constant, perhaps indicating perpetuation of a particular population of cells. These cells perhaps reflect their dependency on glycolysis rather than oxidative phosphorylation as their primary source of energy, hence resistant to tigecycline treatment (Farge *et al.*, 2017). These cell population upon characterization with flow cytometry indicated an increase in c-kit expression, in contrast to the observation in *MLL-AF9* cells where no changes in c-kit expression was observed. This increase in c-kit expression in *Kat2a* WT cells undergoing transformation with *RUNX1-RUNX1T1(9a)* suggested an accumulation of immature myeloid progenitor cells, which was contrary to the observations in *MLL-AF9* cells where an enhanced differentiation was observed upon tigecycline treatment, characterized by an increase in proportion of non-compact colonies. Although both the models highlight contrasting downstream effects upon inhibition of mitochondrial translation using Tigecycline,

these observations may justify how *Kat2a* mediated impairment in mitochondrial biogenesis pathway associated genes may cause impairment in leukaemia stem-like cells in case of *MLL-AF9*, whereas it leads to perpetuation of leukaemia initiating cells and further accelerate the disease progression in case of *RUNX1-RUNX1T1(9a)*, suggesting a model-specific role.

Overall, the analysis has an implication towards understanding the association of mitoribosomes in human pathologies using *Kat2a* perturbation. The defects in mitochondrial translational machinery can arise from mutations in the components of machinery, including tRNAs, aminoacyl-tRNA synthetases, translation factors, and ribosomal components (Vafai and Mootha, 2012; Boczonadi and Horvath, 2014), some of which have been documented before (Saada *et al.*, 2007; Galmiche *et al.*, 2011; Smits *et al.*, 2011; Carroll *et al.*, 2013). These mutations commonly cause instability of the protein, impaired assembly of the affected mitoribosomal subunit, a deficiency in oxidative phosphorylation, and a variety of severe phenotypes, including dysmorphism, lactic acidosis, neurological disorders, and cardiomyopathies, which are often fatal early in life.

Since most of these proteins are involved in apoptotic signalling and the regulation of cell proliferation, their altered expression has been associated with the development of cancer (Greber and Ban, 2016; Kampen *et al.*, 2020). Further to this, as indicated from the above experiments, elevated mitochondrial translation may be needed to provide the metabolic capacity to meet the energy requirements of cancer cells. Even though ATP production in cancer cells has long been thought to occur mostly by glycolysis in the cytoplasm (Hanahan and Weinberg, 2011), recent evidence has indicated that some types of cancer cells may rely heavily on oxidative phosphorylation in their mitochondria, utilizing metabolites provided by neighbouring glycolytic stromal cells (Pavlides *et al.*, 2009) or by cancer cells in more poorly oxygenated regions of the cancer (Feron, 2009). In agreement with these ideas, the upregulation of a large number of mitoribosomal proteins has been observed in human breast cancer cells, but not in adjacent stromal cells, leading to the proposal that mitochondria fuel epithelial cancer cell metabolism (Sotgia *et al.*, 2012). Therefore, targeting mitochondrial translation may be a promising strategy for cancer therapy (Lamb *et al.*, 2015). Interfering with mitochondrial translation by treatment with tigecycline or the knock down of mitochondrial EFTu selectively inhibits the proliferation of leukaemia cells, compatible with insights from *MLL-AF9* model,

suggesting that tigecycline may possibly be an anticancer agent (Järås and Ebert, 2011; Škrtić *et al.*, 2011).

7.3 *Kat2a* loss may serve as a model to study ribosomopathies

The *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation in *Kat2a* NULL cells also captured transcriptional instability in ribosomal biogenesis and cytoplasmic translation associated gene expression programmes. The consistent capture of these gene expression pathways in different comparisons during pre-leukaemia transformation as well as in the absence of *Kat2a* are compatible with insights from literature. Previous studies have shown that ribosomal biogenesis is corrupted during pre-leukaemia transformation in *RUNX1* mutant haematopoietic Stem Progenitor cells (HSPCs) which correlated with lower rates of translation and a stress resistant phenotype, thus providing them with a selective advantage over normal HSCs in the bone marrow (Cai, Gao, Teng, Ge, *et al.*, 2015). Further, studies have indicated that *RUNX1-RUNX1T1* binds to ribosomal DNA sequences and epigenetically regulates pre-rRNA synthesis (Bakshi *et al.*, 2008). However, *Kat2a* associated impact on ribosomal biogenesis may be through Myc-dependent regulation of rRNA genes. Myc activates co-regulatory proteins, such as the transformation/transcription domain-associated protein (TRRAP) by binding to the promoter. TRRAP is part of the histone acetyltransferase (HAT) complex, which is responsible for targeting acetylation of histones through the acetyltransferase activity of the *Kat2a*, thereby impacting the transcription of rRNA genes (Gomez-Roman *et al.*, 2003; Van Riggelen, Yetil and Felsher, 2010). Since, *Kat2a* directly targets Myc as shown by previous studies in our lab (Domingues *et al.*, 2020), future studies may explore whether Myc dependent regulation of rRNA genes is via Kat2a mediated acetylation.

On these lines, the impact of *Kat2a* loss was studied on protein synthesis rate using OP-Puro incorporation method (Hidalgo San Jose and Signer, 2019) where a significant reduction in OP-Puro high (Hi) cell population was observed within the $Lin^{-}c\text{-}kit^{+}$ compartment of cells in *Kat2a* NULL cells transformed with *Idh1*R132H mutation. This was also accompanied by a gain in OP-Puro low (Lo) cell population, suggesting that loss of *Kat2a* reduces translation rate in *Idh1*R132H pre-leukaemia cells. These observations were compatible with previous studies conducted in our lab on *MLL-AF9* model where a reduction in ribosomal biosynthetic

programmes was suggested, based on scRNA-seq analysis in *MLL-AF9* transformed *Kat2a* NULL cells. Upon performing OP-Puro incorporation assay in a subset of leukaemia cells defined by $Lin^{-}Kit^{+}Sca1^{-}CD34^{+}FC\gamma R^{+}$ (L-GMPs), a reduction in protein synthesis was observed in *Kat2a* NULL cells (Domingues *et al.*, 2020), in line with the insights from *Idh1*R132H pre-leukaemia cells. Altogether, the analysis suggested that depletion in ribosomal biosynthetic programmes can impact *Kat2a* mediated pre-leukaemia progression by providing a selective advantage to the transformed cells over normal HSCs in the bone marrow, compatible with the observation made by Speck and colleagues (Cai, Gao, Teng, Mason, *et al.*, 2015).

The mechanistic association of impairment of ribosomal biosynthetic programmes during pre-leukaemia transformation was confirmed using S6K1 inhibitor which acts on S6K1 isoform of p70 ribosomal S6 kinase and thus inhibits protein synthesis (Pearce *et al.*, 2010). S6K1 is a member of the A, G, and C family of serine/threonine protein kinases and is a key regulator of cellular metabolism. S6K1-deficient mice are smaller than wild-type littermates and display hypoinsulinemia and glucose intolerance (SH, D and G, 2006). To mediate its effects on metabolic pathways, S6K1 phosphorylates a number of downstream substrates including the small ribosomal subunit protein S6 (rpS6) (Salmond *et al.*, 2015). *Kat2a* WT cells undergoing transformation with *RUNX1-RUNX1T1(9a)* showed an enhanced colony forming potential at plating 2 upon S6K1 inhibition, compatible with the insights from scRNA-seq analysis suggesting that impairment of ribosomal biogenesis/cytoplasmic translation may contribute towards the acceleration in pre-leukaemia transformation. This increase in colony forming potential was lost markedly upon plating 3 indicating that inhibition of cytoplasmic translation is beneficial for *RUNX1-RUNX1T1(9a)* pre-leukaemia progression during early stages of transformation. Similar observations were made for *Idh1*R132H transformed *Kat2a* WT cells indicating that inhibition of cytoplasmic translation is beneficial for *RUNX1-RUNX1T1(9a)* and *Idh1*R132H pre-leukaemia progression during early stages of transformation.

The analysis was further extended to study association with *MLL-AF9* transformed cells based on the scRNA-seq data on *MLL-AF9* leukaemias revealing ribosomal biogenesis/cytoplasmic translation amongst the top enriched categories which were found to be downregulated upon *Kat2a* loss (Domingues *et al.*, 2020). The *MLL-AF9* transformed *Kat2a* WT cells generated *in*

vitro upon S6K1 inhibition showed a reduction in colony forming potential, in line with the observation in *MLL-AF9* model where loss of *Kat2a* aids in depletion of leukaemia stem-like cells (Domingues *et al.*, 2020). Upon further characterization of colonies, a significant reduction in compact colonies, representative of immature myeloid cells, was observed which was complemented by a gain in mixed and dispersed colonies representing differentiated macrophages. Overall, the S6K1 inhibition study showed contrasting dependencies on ribosomal biosynthetic programmes in *RUNX1-RUNX1T1(9a)* and *Idh1*R132H pre-leukaemia models compared to *MLL-AF9* maintenance model, consistent with the insights drawn from mitochondrial bioenergetics associated mechanistic experiments. The analysis suggested that early reduction in expression of genes associated with ribosomal biosynthetic programmes associated may contribute towards perpetuation of leukaemia initiating cells and further accelerate disease progression in *RUNX1-RUNX1T1(9a).* However, once the leukaemia is established, inhibition of ribosomal biogenesis assembly promotes cellular differentiation leading to delayed leukaemogenesis as observed in *MLL-AF9* leukaemia, suggesting the stage-specific role of ribosomal biosynthetic programmes during leukaemia progression.

The mechanistic association of impairment in ribosomal biosynthetic programmes during *RUNX1-RUNX1T1(9a)* and *Idh1*R132H leukaemia initiation was consistent with the defects in ribosome biogenesis observed in different haematological malignancies including, Diamond-Blackfan anaemia (DBA). DBA is a congenital syndrome associated with anaemia, physical malformations, and cancer (Vlachos *et al.*, 2012b). In the majority of individuals with DBA, mutations or gene deletions of a subset of ribosomal proteins are found, with RPS19 mutations accounting for about 25% of all cases (Flygare *et al.*, 2007; Devlin *et al.*, 2010). The major mechanism of DBA disease progression is associated with haploinsufficiency of a ribosomal protein that disrupts the processing of pre-ribosomal RNA (pre-rRNA), leading to abortive ribosome biogenesis (Choesmel *et al.*, 2007; Farrar *et al.*, 2014). Individuals with DBA have an increased frequency of both solid cancers and leukaemia (Vlachos *et al.*, 2012b), where specifically RPL5 or RPL11 loss-of-function have been observed to predispose DBA patients to cancer by loss of capacity to activate the TP53 tumour suppressor (Molavi, Samadi and Hosseingholi, 2019). DBA shows an enhanced erythroid defect which can be attributed to mutations in RPS19 and RPL11 which reduce Internal Ribosome Entry Site (IRES)-mediated translation of erythroid differentiation factors BAG1 and CSDE1 in DBA both in mouse

models and patient samples (Horos *et al.*, 2012). In contrast to this, impairment of ribosomal biogenesis upon *Kat2a* loss did not show defects associated with any haematopoietic compartment including in *RUNX1-RUNX1T1(9a)*, *Idh1*R132H, and *MLL-AF9* models of leukaemia (Domingues *et al.*, 2020), suggesting that the observations might be disease-specific.

Somatic recurrent ribosomal protein mutations and deletions have also been detected in T-cell acute lymphoblastic leukaemia (T-ALL), where low frequency mutations in RPL11 are observed (Tzoneva *et al.*, 2013b). In addition to this, heterozygous loss of RPL5 by deletions or inactivating mutations has been shown to occur in 2% of T-ALL samples whereas mutations or deletions in RPL22 were described in 4% of T-ALL samples indicating the disease progression to be dependent on reduced ribosomal activity (De Keersmaecker *et al.*, 2013b; Liu *et al.*, 2017). 20% of the relapsed chronic lymphoid leukaemia (CLL) patients show mutations in RPS15 thereby suggesting that reduced levels of mature ribosomes and lower overall translation weaken the cells and promote a quiescent state (Ljungström *et al.*, 2016b; Bretones *et al.*, 2018). These observations are overall compatible with the reduction in ribosomal activity in *RUNX1-RUNX1T1(9a)* pre-leukaemia upon *Kat2a* loss and indicate that *Kat2a* loss may serve as a model to study ribosomopathies as a cause of leukaemia transformation.

7.4 Loss of *Kat2a* promotes cellular diversification with an increase in transcriptional variability

Pre-leukaemia cells serve as reservoir of early haematopoietic multipotent progenitors which are capable of diversifying during the process of leukaemia development. This cellular diversification usually occurs in a discontinuous manner which further gives rise to discrete developmental stages, including progenitor and differentiated states, as well as discrete lineages and terminally differentiated types (Koeffler and Leong, 2017). This process of cell dynamics is somewhat similar to the stages of development and was first noted by C. Waddington in the 1940s. His landscape describes an iconic picture where a marble rolls down a surface, staying in valleys and seeking the lowest point. At watersheds, the valleys branch so that the marble takes one of two available paths. In Waddington's picture, the ball represents a

developing cell in an embryo and the landscape epitomizes some more abstract set of constraints, thus clearly heralding the notions of stability and instability in the modern sense of dynamics (Wang *et al.*, 2011). Indeed, it has recently become clear that Waddington's epigenetic landscape in principle represents the dynamics of a system of gene regulatory interactions that impose constraints to and drive cell development (Huang, 2009; MacArthur, Maayan and Lemischka, 2009), giving a metaphor for the qualitative understanding of developmental processes of cells. However, similar understanding on how the dynamics of a gene regulatory circuit governs binary cell fate decisions aiding in pre-leukaemia to leukaemia progression is yet to be explored.

The phenomenon of cellular diversification is an independent and reproducible phenomenon and has been seen in case of various developmental systems including embryonic and haematopoietic systems (Kamminga *et al.*, 2000; Keller, 2005; Lüer and Technau, 2009). The consequent process of cell fate specification highlights the relationship between cell lineages and genetic programmes and can be defined as a sequence of instructions associated with a given fate, specific decision events, directionality, and a means of apportioning a defined numbers of cells to particular fates to generate proportionate tissues. Herein, the mammalian homolog of Gcn5, known as *Kat2a,* has been utilized as a tool to study binary cell fate decisions contributing towards pre-leukaemia to leukaemia progression. *Kat2a* is a well-established regulator of stochastic transcriptional regulation (Raser and O'Shea, 2004b), the deletion of which results in mutants possessing enhanced cell-to-cell transcriptional variability in gene expression as measured across a range of locus fluorescence reporters (Weinberger *et al.*, 2012b). The cell-to-cell transcriptional variability reflects the variability in the number of mRNA molecules produced from a given locus at a certain time point (Ko, 1991; Mcadams and Arkin, 1997). The same can be measured on the basis of snapshot analysis performed on gene expression studies (Sanchez, Choubey and Kondev, 2013). One of the major causes behind the phenomenon of cell-to-cell gene expression variability is the occurrence of transcription in the form of 'bursts' (Cai, Friedman and Xie, 2006). mRNA is produced from the promoter of a gene in the form of bursts, switching between periods of prolonged inactive (or 'off') states and short-lived active (or 'on') states (Thattai and Van Oudenaarden, 2001). Most gene loci reflect the phenomenon of transcriptional bursting with a characteristic size and frequency, where size of the burst indicates the number of transcript molecules produced during

individual bursts and burst frequency reflects the rate at which the promoter is engaged in active transcription (Peccoud and Ycart, 1995; Kumar, Singh and Kulkarni, 2015). These rates of binding and dissociation are majorly dependent on protein availability and the presence of other bound proteins. Both burst size and burst frequency contribute to mean gene expression, whereas transcriptional variability is more strictly dependent and shown to be anti-correlated with burst frequency (Hornung *et al.*, 2012a). The size and frequency of bursts in yeast are increased through histone acetylation of genes and promoters, respectively (Weinberger *et al.*, 2012b). In this study, the presence of transcriptional variability was assessed during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation upon *Kat2a* loss, based on a Spearman's correlation coefficient measuring cell-to-cell variability.

The global comparison indicated an increase in transcriptional variability upon loss of *Kat2a* at respective time points. The increase in transcriptional variability upon *Kat2a* loss was more evident at 2 months compared to 4 months, compatible with the cellular dynamics observed at early stages of pre-leukaemia transformation. Further, there was a reduction in transcriptional variability at 4 months compared to 2 months in both the genotypes, perhaps indicating that increase in cell-to-cell transcriptional variability is the hallmark of an early pre-leukaemia transformation event. The increase in transcriptional variability was compatible with the previous laboratory work on *MLL-AF9* model of leukaemia where a significant increase in gene expression variability was measured by increase in coefficient of variation upon *Kat2a* loss. This increase in transcriptional variability in *MLL-AF9* transformed *Kat2a* knockout leukaemias was evident at all levels of gene expression and associated with greater cell-to-cell dispersion in transcript levels (Domingues *et al.*, 2020). The variability in transcript levels was attributed to a change in burst frequency but not in burst size, in line with the evident role of yeast orthologue, Gcn5 (Weinberger *et al.*, 2012b).

Cell-to-cell variation in gene expression and fluctuations over time in single cells have broad implications. This noisy gene expression is often perceived as undesirable and unpredictable; however, living systems are inherently noisy and are optimized to function in the presence of stochastic fluctuations (McAdams and Arkin, 1999). The variability in gene expression is sometimes considered harmful as it distorts cell signals, corrupts circadian clocks, and disrupts the fine-tuned process of step-wise development (Barkai and Leibler, 2000; Paulsson, Berg and

Ehrenberg, 2000). On the other hand, some organisms can exploit stochasticity to introduce diversity into a population, as occurs with the lysis–lysogeny bifurcation in phage λ (Robb and Shahrezaei, 2014) or the DNA inversion mechanism in bacteria (Van de Putte and Goosen, 1992). The variability in gene expression has also been directly implicated as a mechanism of cell fate choice in yeast (Blake *et al.*, 2006b) and has been previously shown to be associated with normal transitions into haematopoietic lineage specification (Pina *et al.*, 2012; Teles *et al.*, 2013).

The cell-fate decisions and lineage dependencies upon *Kat2a* loss were studied during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation using pseudotemporal ordering. Pseudotemporal ordering of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT cells resulted in a linear trajectory, starting from a cell population marked by c-Kit^{+}Ly6e^{hi}Cd34^{+}Flt3^{+}, representing lymphoid primed multipotential progenitor candidates (LMPP) with a subset of cells highlighted by c-Kit^{+}Ly6e^{hi}Cd34^{+}Flt3^{-}, representing the presence of multipotent progenitors (MPP). These populations were identified based on previous literature insights (Weissman and Shizuru, 2008)(Doulatov *et al.*, 2012). This cell population characterized by LMPP-like and MPP was enriched in *Kat2a* WT cells collected at 2 months post transplantation, suggesting that these cells are still at the early stages of haematopoietic hierarchy and haven't yet progressed to pre-leukaemia transformation. These multipotent progenitors then showed a bifurcation, one arm of which was towards Granulocyte-Macrophage Progenitor (GMP)-like population highlighted by Ly6e^{lo}Cd34^{+}Fcgr3^{+}, whereas the other arm showed the presence of monocyte-macrophage population characterized by Cd14^{+}Cd74^{+}Mafb^{+}Mafg^{+}. The GMP-like population was relatively more enriched in *Kat2a* WT cells collected 4 months post transplantation confirming the pre-leukaemia hierarchical progression of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT cells. The monocyte-macrophage arm of the trajectory included *Kat2a* WT cells processed 2 months post transplantation suggesting that differentiation of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* WT cells towards monocytic lineage is an early event during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation.

Similarly, pseudotemporal ordering of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells was performed which yielded a relatively branched trajectory. This was in contrast to the trajectory followed by *Kat2a* WT cells, suggesting that loss of *Kat2a* promotes heterogeneous

transcriptional profiles which further leads to varied cellular trajectories during the process of *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation. The different subsets of cell population present in the *Kat2a* NULL trajectory overlaid with the cellular population identified in the trajectory pursued by *Kat2a* WT cells, including the presence of LMPP-like and MPP-like cells. However, I observed emergence of three different branches from LMPP-like cell population, two of which coincided with the *Kat2a* WT trajectory and were identified as GMP-like and monocyte-macrophage-like. In this case, however, the proportion of cells belonging to monocyte-macrophage-like cell population were higher compared to *Kat2a* WT cells, indicating that loss of *Kat2a* may promote differentiation towards monocytic lineage during the process of *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation.

Overall, the analysis captured different compartments of haematopoietic hierarchy which led towards accumulation of *RUNX1-RUNX1T1*$^{+}$ pre-leukaemic progenitor cells, in line with the insights obtained from *in vivo* analysis of *RUNX1-RUNX1T1(9a)* pre-leukaemia cells, where an accumulation of early myeloid progenitors denoted by GFP^{+}c-Kit^{+}FcγR^{+} was observed. The global trajectory analysis indicated a cell population marked by c-Kit^{+}Ly6e^{hi}Cd34^{+}Flt3^{+}, representing lymphoid primed multipotential progenitor candidates (LMPP) with a subset of cells marked by c-Kit^{+}Ly6e^{hi}Cd34^{+}Flt3^{-}, representing the presence of multipotent progenitors (MPP). These multipotent progenitors then showed a bifurcation, one arm of which was towards Granulocyte-Macrophage Progenitor (GMP)-like population marked by Ly6e^{lo}Cd34^{+}Fcgr3^{+}, whereas the other arm showed the presence of monocyte-macrophage population as well as B-cell lymphocyte progenitors characterized by Cd14^{+}Cd74^{+}Mafb^{+}Mafg^{+} and Cd79a^{+}Cd19^{+}Il7r^{+}, respectively. The monocyte-macrophage arm along with B-cell lymphocyte progenitors present in the trajectory were enriched in *Kat2a* NULL cells processed 2 months post transplantation suggesting that differentiation of *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells towards monocytic and B-lymphocytic lineage is an early event during *RUNX1-RUNX1T1(9a)* pre-leukaemia transformation. The hierarchical organization of trajectory from LMPP to GMP coincided with the changes in cell-to-cell transcriptional variability with a marked reduction from LMPP-like to GMP. However, consistent with global observations, an increase in gene expression variability was observed in individual cell compartments upon loss of *Kat2a* at each time point.

The early loss of *Kat2a* indicated cellular differentiation towards B-cell and monocytic lineages suggesting that loss of *Kat2a* is capable of diversifying the pre-leukaemic population of cells. These observations were experimentally validated by utilising an *in vitro* set-up where both *Kat2a* WT and *Kat2a* NULL cells were transduced with *RUNX1-RUNX1T1(9a)* and the levels of B220, a B-cell marker, were assessed using flow cytometry. The analysis indicated an increase in B-cell population upon *Kat2a* loss. A similar observation for F4/80, a monocyte marker, was made, where an increase in F4/80 expression was observed in *Kat2a* NULL cells undergoing transformation with *RUNX1-RUNX1T1(9a).* Strikingly, none of these observations held true in further re-platings, confirming that the differentiation towards B-cell and monocytic lineage is an early effect seen upon loss of *Kat2a*. However, it is worth noting that these observations may be consequent to the fact that these cells were maintained in myeloid culture conditions and continued serial replating selects for the cells with enhanced self-renewal potential. The presence of monocytic differentiation upon *Kat2a* loss in *RUNX1-RUNX1T1(9a)* pre-leukaemia was consistent with the previous work conducted in our lab where *MLL-AF9* transformed cells in absence of *Kat2a* also showed an enhanced monocytic differentiation as marked by an enhanced expression of Cd11b monocytic marker (Domingues *et al.*, 2020). The differentiation towards B-cell lymphocytic lineage was also in line with the literature insights pointing towards an aberrant expression of B-cell markers, like Cd19 and Cd79a, due to constituent expression of B-cell transcription factor PAX5 in t(8;21) patient-derived cells, which binds to the promoter and enhancer of some of these genes and thus contributes towards their aberrant expression through epigenetic regulation (Walter *et al.*, 2010) (Kostareli *et al.*, 2012). Moreover, the yeast homologue of *Kat2a*, known as *Gcn5*, has been found to participate in transcriptional regulation of the IgM H-chain gene via histone acetylation in immature B-cells, suggesting it's key role in epigenetic regulation of B cell functions (Kikuchi *et al.*, 2014b, 2014a). Overall, these insights were consistent with our observations in *MLL-AF9* leukaemia where *Kat2a* depletion enhanced the probability of differentiation cell fate transitions (Domingues *et al.*, 2020). This was also akin to previous lab observations on depletion of *Kat2a* in mouse embryonic stem cells where an increase in gene expression variability was negatively correlated with stabilization of pluripotency (Moris *et al.*, 2018b).

Upon following the pre-leukaemia transformation hierarchy based on the progression of myeloid arm of bifurcation, pre-leukaemia progenitors were identified marked by c-Kit^{+}Ly6e^{-} Fcgr3^{+}, compatible with the characterization of pre-leukaemia *in vivo.* These cells were majorly enriched in *Kat2a* WT and *Kat2a* NULL cells studied at 4 months post-transplantation, thus highlighting an accumulation with respect to time. The gradual building-up of leukaemia progenitors with time coincided with the progressive *RUNX1-RUNX1T1(9a)* transformation. The increased accumulation of leukaemia progenitors upon *Kat2a* loss at respective time points was in line with the *in vivo* observation of accelerated leukaemia progression upon loss of *Kat2a*. These cells upon further characterization were found to have higher average expression of *RUNX1-RUNX1T1* (Ptasinska *et al.*, 2012) which were exclusively present in this compartment of cells, thus suggesting the presence of leukaemia-associated events. Furthermore, a gradual reduction in *Kat2a* targets was observed in the leukaemia progenitor cells relative to LMPP-like and GMP population highlighting *Kat2a* mediated events in this compartment of cells. Accordingly, the progression from GMP to pre-leukaemia was marked by an increase in gene expression variability, consistent with the presence of leukaemia associated events and in line with our previous observations in *MLL-AF9* model of leukaemia (Domingues *et al.*, 2020). Altogether, the trajectory analysis suggested the propensity of early progenitors to diverge towards *RUNX1-RUNX1T1(9a)* mediated transformation upon *Kat2a* loss, as evident from the accumulation of pre-leukaemia progenitors.

One of the ways by which cells can change their state is by stochastic fluctuations that allow them to jump around the landscape without requiring any parameter changes (Moris, Pina and Arias, 2016b). These stochastic fluctuations are an intrinsic feature of some biological systems, especially with respect to transcription in which low numbers of transcription factors and DNA molecules can cause sporadic engagement of RNA polymerase and lead to discontinuous mRNA synthesis (Raj and van Oudenaarden, 2008) (Eldar and Elowitz, 2010) resulting in deviations from the stable cell state. Usually, these perturbations are small enough allowing the system to be back to its original stable state. On the other hand, these events could represent stochastic fluctuations of transcription factor levels over a threshold that proves sufficient to promote cell identity changes (Schröter *et al.*, 2015). Such events have been observed in bacteria, in which the stochasticity is shown to have functional significance (Süel *et al.*, 2006).

Stochastic fluctuations might also apply to embryonic stem cell populations and in general in eukaryotic stem cells (Arias and Hayward, 2006; Martinez Arias and Brickman, 2011).

Strikingly, the proportion of cells belonging to GMP compartment including both *Kat2a* WT and *Kat2a* NULL population at 2 months was lower than the combined *Kat2a* WT and *Kat2a* NULL population of cells at 4 months pointing towards an abrupt transformation trajectory owing to the enhanced transcriptional variability observed upon trajectory analysis. This observation suggested that gene expression variability during pre-leukaemia transformation need not align through time along an ordered continuum and could instead be interpreted as a reflection of systemic features that allow a cell to stochastically rearrange its networks. It may also suggest that the actual transitions between states might be discontinuous, and that the observed cell-to-cell transcriptional variability in pre-leukaemia compartment may reflect the existence of a dynamic array of transcriptional states, with varied probability of transitions. Previous literature insights suggest that discontinuous fate decisions may occur in few systems including haematopoietic system in the case of early myelo-erythroid choices in mouse haematopoiesis and in human haematopoietic stem cells that can give rise to differentiated cells directly without going through multilineage intermediates (Pina *et al.*, 2012; Notta *et al.*, 2016). In fact, discontinuity in cell-fate trajectories has been observed upon loss of *Kat2a* in *MLL-AF9* leukaemia where *Kat2a* WT *MLL-AF9* cells aligned along an almost linear trajectory. This was in contrast to *Kat2a* knockout *MLL-AF9* leukaemia cells which were distributed along multiple differentiation trajectories. These multiple trajectories reflected multiple uncoordinated routes into cell fate decision-making, which were initiated but not coherently completed by cells depleted of stem cell potential (Domingues *et al.*, 2020). Similar insights were obtained upon *Kat2a* depletion in mouse embryonic stem cells where *Kat2a*-inhibited cells were incapable of sustaining pluripotency, but lagged in their ability to proceed through lineage differentiation (Moris *et al.*, 2018b).

Altogether, the insights generated from lineage dependency studies suggested that loss of *Kat2a* enhances the propensity of cell fate transitions which results in increased molecular heterogeneity and stochastic fate choices. *Kat2a* loss likely facilitates rather than determines fate acquisition by increasing cell-to-cell gene expression variability and may function similarly in the context of other malignancies. Future studies may involve a deeper insight into

understanding the gene regulatory modules associated with individual cell state and during the process of diversification and may aid in a better molecular understanding of cell fate dynamics induced upon *Kat2a* loss during pre-leukaemia.

7.5 Increase in transcriptional variability upon *Kat2a* loss may be consequential to differential chromatin accessibility

The process of cell-fate decision making requires changes in transcriptional state of a cell. The transcription of a gene requires interaction with histone modifiers and chromatin remodellers, which further determine the accessibility of the transcription factors to the DNA and their binding stability (Moris, Pina and Arias, 2016b). The alterations in chromatin accessibility may act as a key to changing the coordinates of the transcriptional state of a cell (Swain, Elowitz and Siggia, 2002; Chalancon *et al.*, 2012). For example, the binding of transcription factors can define cellular state by activating specific genes and thus define which gene regulatory modules will be active. To associate gene expression variability observed upon *Kat2a* loss with alteration in chromatin accessibility patterns, single-cell ATAC sequencing was performed. Single-cell Assay for Transposase Accessible Chromatin-sequencing or scATAC-seq is a method which employs Tn5 transposase, which simultaneously tags and fragments DNA sequences in open chromatin regions, and reveals the interplay between genomic locations of open chromatin, DNA-binding proteins, individual nucleosomes and chromatin compaction at single-cell resolution (Buenrostro *et al.*, 2015b). The assay was performed on Kasumi-1 cells representative of the same primary genetic alteration, *RUNX1-RUNX1T1(9a).*

Upon performing differential accessibility analysis using fisher exact test and information gain methods (Baker *et al.*, 2019), 50 peaks were characterized as being unique to DMSO, 520 peaks unique to MB-3 and 3587 peaks which were attributed as common set of peaks. Strikingly, within the common subset of peaks, no significant differences were observed in terms of differential accessibility upon inhibition of KAT2A suggesting that attenuating KAT2A activity does not impact global chromatin accessibility in Kasumi-1 cells. This was unsurprising given the acetyltransferase activity of Kat2a and capability to act as a chromatin remodeller by stabilising transcription rather than initiating it (Jin *et al.*, 2014). Although, this observation may correlate with the fact that no changes in transcriptional variability were

observed in *RUNX1-RUNX1T1(9a)* transformed *Kat2a* NULL cells at 4 months relative to *Kat2a* WT cells within the leukaemia progenitor compartment obtained from scRNA-seq data, likely due to establishment of leukaemia by that time point of sampling. However, unpublished analysis in the Pina lab using scRNA-seq of MB-3 treated Kasumi-1 cells highlighted enhanced cell-to-cell transcriptional variability upon inhibition of KAT2A. These findings overall indicated that increase in transcriptional variability cannot be solely attributed to differential chromatin accessibility patterns.

The analysis conducted on understanding the region-gene associations suggested an overall rearrangement in the distribution of peaks upon KAT2A inhibition, based on their vicinity with TSS. In control samples treated with DMSO, 41.66% of the peaks were found to be proximal to TSS whereas in case of MB-3 treated cells, 37.68% of the peaks were proximal to TSS. However, an enhanced chromatin accessibility was observed at distal peaks upon KAT2A inhibition where 38.69% of the distal peaks were annotated compared to 30.55% of the peaks in control cells. This may indicate a crosstalk between the proximal and distal regulatory regions speculating about higher order chromatin organization of KAT2A associated targets. This was compatible with the previous observations in our lab where *Kat2a* loss in primary *MLL-AF9* cells was found to be associated with mild increase in H3K9ac, a mark deposited by *Kat2a*, at candidate active enhancer regions. These candidate enhancer regions were identified by the presence of H3K27ac, suggesting a possible pattern of imbalance of H3K9ac regulation between promoters and enhancers (Domingues *et al.*, 2020).

Going forward, it will be interesting to develop an understanding towards this promoter vs enhancer specificity observed upon inhibition of KAT2A, in light of the enhancer-promoter interaction via CTCF binding. CTCF gene encodes a transcriptional regulator protein with 11 highly conserved zinc finger domains. The using of different combination of eleven ZF domains allows this protein to bind different DNA sequence and interact with various protein factors (Ohlsson, Renkawitz and Lobanenkov, 2001) (Ren and Zhao, 2019). A recent study has suggested that there is a conserved subset of CTCF sites using prostate cancer as a model system, which are resistant to CTCF depletion. The authors have further proposed that these persistent CTCF sites are essential for constitutive higher order chromatin architecture and the maintenance of long-range epigenetically regulated domains (Khoury *et al.*, 2020).

CTCF binding sites are present widely in genome and a majority of them are located within Topologically Associated Domains (TADs), defined as megabase-sized local chromatin interaction domains (Nora *et al.*, 2020). these intra-domain CTCF binding sites are in the close vicinity of potential enhancers of transcription, as marked by p300 and H3K4me1, and thus may influence the activity of enhancers (Ren and Zhao, 2019). A study performed on utilising chromosome conformation capture carbon copy (5C) technology in human cell lines specifically, GM12878, K562 and HeLa-S3 cells, found that a fraction of CTCF enriched distal elements significantly interact with gene promoters. These findings suggested that one of the main roles of CTCF in genome function may be to facilitate the interaction between regulatory sequences and promoters (Sanyal *et al.*, 2012). Since distal enhancers must physically contact with their target promoters to carry out their activity, the nearby CTCF molecules may bring enhancers to the vicinity of their target promoters (Ren *et al.*, 2017b; Nora *et al.*, 2020). CTCF can mediate the enhancer-promoter contact through the interaction between CTCF bound nearby enhancers and cohesin loaded nearby promoters (Fudenberg *et al.*, 2016; Merkenschlager and Nora, 2016). Liu and colleagues reported that regulatory elements-bound CTCF and cohesin can recruit the core promoter factor TAF3 and mediate its contact with promoters through TAF3-dependent loop formation in ES cells. They further showed that depletion of CTCF reduces the efficient recruitment of TAF3 to distal regulatory elements, compromises endoderm differentiation marker gene expression, such as Gata4, Afp, and Apoa1(Liu *et al.*, 2011). Another example from a relatively recent study highlighted the interaction between CTCF and enzyme poly-ADP-ribose (PARP1), which allows establishment of inter-chromosomal contacts between active circadian loci and repressive chromatin at the lamina, thereby mediates circadian transcriptional plasticity (Zhao *et al.*, 2015). These studies highlighted the role of promoter-enhancer interaction facilitated by CTCF in changing the transcriptional landscape of cells.

As discussed earlier, recent studies have indicated the role of promoter sequence, nucleosome positioning, epigenetic modifications, and three-dimensional genome organization contributing towards regulation of gene expression (Chalancon *et al.*, 2012). On similar lines, gene expression variation may result from any fluctuation of above-mentioned factors, especially for CTCF mediated promoter-enhancer interaction. A recent study has associated

gene expression variability with CTCF mediated interactions where CTCF-bound T cell-specific genes GATA3, CD90, CD28, CD5 displayed significantly increased expression variation in CTCF depleted cells (Ren *et al.*, 2017b). However, the increased cell-to-cell variation of expression by knocking down of CTCF could also be accounted for by the heterogeneous CTCF knockdown efficiency across different cells. Conclusive evidence came from the deletion of a specific CTCF binding site at Thy1 locus, nearby a distal enhancer, using CRISPR/CAS9, which resulted in a significantly higher cell-to-cell variation of gene expression in the CRISPR knockout cells. The enhanced variability in gene expression by deletion of CTCF binding site at Thy1 locus was correlated with decreased Thy1 promoter-enhancer interaction but not changes in the TAD structure, strongly suggesting a model that CTCF binding near the enhancer region stabilizes the interaction between the Thy1 promoter and its enhancers and thus reduces the cell-to-cell variation of Thy1 expression (Ren *et al.*, 2017b). These observations were compatible with the increase in cell-to-cell transcriptional variability observed in both *MLL-AF9* and *RUNX1-RUNX1T1(9a)* models of leukaemia. These insights hold more relevance as H3K9ac-depleted promoters in *Kat2a* knockout *MLL-AF9* leukaemia cells had a significant association with CTCF binding leading to speculation that CTCF may be dislodged to enhancers and promote asymmetric distribution of histone acetylation marks, with dysregulation of locus control.

Although inhibition of KAT2A did not reveal any global changes in chromatin architecture in Kasumi-1 cells, upon segregating the cells based on dimensionality reduction algorithm and further k-medoid clustering, the KAT2A inhibited cells corresponded to two different clusters. The presence of two different clusters upon KAT2A inhibition pointed towards the presence of differential peak footprint which may be consequential to the rearrangement in the distribution of peaks which were observed upon KAT2A inhibition. The peaks segregating these two clusters upon MB-3 treatment contributed towards cell cycle specifically G2/M transition as well as monocytic differentiation, suggesting diversification of cellular states upon KAT2A inhibition. This was compatible with insights from scRNA-seq obtained from *RUNX1-RUNX1T1(9a)* pre-leukaemia. The higher accessibility peaks obtained upon comparing individual cluster with respect to control cells were biologically interpreted as the genes contributing to cell cycle, DNA repair, epigenetic modifications, and protein ubiquitination, in line with some of the categories enriched during *RUNX1-RUNX1T1(9a)* pre-leukaemia

transformation. Interestingly, the genes associated with mitochondrial bioenergetics and ribosomal biosynthetic programmes, which were captured in scRNA-seq data generated from *RUNX1-RUNX1T1(9a)* pre-leukaemia, did not show up in scATAC-seq analysis. This could potentially be due to the fact that these are highly expressed genes and its unlikely to attribute small changes in gene expression to gain in chromatin accessibility. The regulation of these genes is likely to be more subtle. Moreover, the capturing of DNA repair associated genes may highlight the incidence of increase in DNA damage during the progression of *RUNX1-RUNX1T1(9a)* leukaemia, in line with the notion that Kat2a/Gcn5 is required for double-strand break DNA repair (Salunkhe *et al.*, 2018), and something which can be explored in detail in future. Although γH2AX staining did not suggest increased DNA damage upon *Kat2a* loss in case of *Idh1*R132H model of pre-leukaemia, this was one of the aspects of studying DNA damage and there may be other mechanisms of inducing DNA damage during pre-leukaemia transformation upon loss of *Kat2a*. The gene expression programmes including protein ubiquitination, cell cycle and epigenetic modifications were also captured consistently in time series comparison conducted using scRNA-seq data individually for *Kat2a* WT cells and *Kat2a* NULL cells during *RUNX1-RUNX1T1(9a)* pre-leukaemia progression and are compatible with literature on gene expression programmes impacted in *RUNX1-RUNX1T1(9a)* leukaemia (Ptasinska *et al.*, 2012, 2014; Martinez-Soria *et al.*, 2018; Nafria *et al.*, 2020). Interestingly, some of these genes, specifically the ones contributing to higher cellular metabolism, were captured during the intermediate stages of *RUNX1-RUNX1T1(9a)* pre-leukaemia hierarchy of progression, potentially linking epigenomic rewiring to stochastic gene expression.

This was further validated by integrating the information obtained from scATAC-seq analysis on Kasumi-1 cells with the scRNA-seq analysis performed on *RUNX1-RUNX1T1(9a)* pre-leukaemia cells. For this, the genes having higher accessibility in one of the MB-3 clusters were utilized for pairwise distance measures using *RUNX1-RUNX1T1(9a)* pre-leukaemia scRNA-seq data. These genes showed an increase in cell-to-cell transcriptional variability in *Kat2a* NULL cells both at 2 months and 4 months with respective *Kat2a* WT cells, consistent with the previous observations from *RUNX1-RUNX1T1(9a)* pre-leukaemia. Further, these genes showed an increase in cell-to-cell transcriptional variability in leukaemia progenitor cells compared to LMPP-like and GMP-like population of cells, in line with the previous scRNA-seq observations. These analyses altogether suggested association between transcriptional

variability observed upon *Kat2a* loss during *RUNX1-RUNX1T1(9a)* pre-leukaemia progression, with differential chromatin accessibility. This association was compatible with previous literature findings where locally disordered methylation arising from stochastic process in primary leukaemia cells was associated with relatively more noisy transcriptional landscape suggesting an association between promoter methylation and gene expression (Landau *et al.*, 2014). Another study conducted by Mason and colleagues showed that the genes containing shifts in epiallele composition at promoters exhibited significantly greater variance in their transcript abundance between relapse and diagnosis, as compared to genes without such epiallele shifts in their promoters. Further, the genes which were differentially expressed in AML relapse patients significantly associated with promoters harbouring epiallele shift than with promoter not harbouring epiallele changes (Li *et al.*, 2016).

However, the scATAC-seq analysis presented here has some caveats associated with small sample size leading to high technical noise. Although stringent filtering of cells might have helped in overcoming some of these, unfortunately, this also means that rare cell subsets having unique chromatin accessibility landscape may have been lost due to strict filtering. It is also worth noting that scATAC-seq data is highly sparse. A typical human scATAC-seq dataset contains 10^2-10^4 cells with 10^3-10^5 sequenced reads per cell. However, the number of cis-regulatory elements (CREs) in the genome far exceeds 10^5 (Ji *et al.*, 2020). Thus, in a typical cell, most CREs do not have any mapped read whereas for the CREs with reads, the number of mapped reads rarely exceeds two because each locus has no more than two copies of assayable chromatin per cell in a diploid genome (Schep *et al.*, 2017; Ji *et al.*, 2020). Also, the current scATAC-seq technology destroys cells during the assay and thus, one can only obtain a snapshot of a cell at one time point. However, molecular events such as transcription factor (TF)-DNA binding and their dissociation are temporal stochastic processes. The steady-state activity of a CRE in a cell is determined by the probability that such stochastic events occur over time. Since probability is a continuous measure, the overall activity of a CRE in a cell should be a continuous signal in principle. The sparse and nearly binary scATAC-seq data collected for each CRE at one single time point therefore cannot accurately describe the CRE's continuous steady-state activity in a cell (Ji *et al.*, 2020). These challenges can be circumvented to certain extent by estimating the probability from static observations using higher cell numbers. The lack of multi-omic technologies capable of monitoring the cells continuously

over time still makes scATAC-seq the most popular method for single-cell regulome mapping due to its relatively simple and robust protocol and unparalleled throughput (Zhou *et al.*, 2019).

Chromatin remodelling has been defined as one of the early candidates responsible for bursting nature of transcription. As eukaryotic genes are wrapped around histone proteins that form chromatin fibres, and chromatin can be remodelled from a tightly bound, transcriptionally inert structure to a more loosely bound, transcriptionally active conformation through the action of various chromatin-remodelling enzymes. Thus, random events of chromatin remodelling could result in random bursts of transcription. Some of the initial studies performed on this hypothesis had provided indications towards the alteration of chromatin-remodelling enzymes resulting in stochastic gene expression (Raser and O'Shea, 2004a) (Xu, Zawadzki and Broach, 2006). Genomic position including positional effects like measuring correlations between proximally located genes also appeared to have a strong effect on covariation in bursting between multiple genes which provided a link between chromatin-related events and stochastic gene expression by indirect means (Becskei, Kaufmann and Van Oudenaarden, 2005) (Raj *et al.*, 2006). In agreement with this, the analysis presented here suggested an association between gene expression stochasticity with a gain in chromatin accessibility pattern during *RUNX1-RUNX1T1(9a)* leukaemia progression.

Recent advancements in single-cell approaches have shown that many eukaryotic genes are transcribed stochastically in bursts of specific sizes and frequencies (Raj *et al.*, 2006) (Levsky *et al.*, 2002) (Larsson *et al.*, 2019), however, very few studies have made an attempt to elucidate the molecular mechanisms behind transcriptional bursting. One such study based on the analysis of 8,000 individual human genomic loci using time-lapse fluorescence microscopy indicated the role of promoter as the key regulator in defining transcriptional dynamics. The study strengthened the argument that transcriptional bursting dominates across the human genome, both burst frequency and burst size vary by chromosomal location, and transcriptional activators alter burst frequency and burst size, depending on the expression level of the locus (Dar *et al.*, 2012). In this line, a study conducted by Naef's lab on histone acetylation using *Bmal1* promoter as their model system suggested that transcriptional bursting predominantly resulted from variations in burst frequency while the genomic position changed the burst size. They further elucidated that promoter histone-acetylation level covaried with burst frequency,

being greatest at peak expression and lowest at trough expression, while remaining unaffected by the genomic location (Nicolas *et al.*, 2018). Although the authors showed it directly for H3K27ac mark however, there was indirect evidence suggesting that H3K9ac would have a similar effect. This was compatible with our previous observation on *Kat2a* associated transcriptional burst frequency in the mouse *MLL-AF9* model where loss of *Kat2a* led to a reduction in burst frequency of *Kat2a*-acetylated targets (Domingues *et al.*, 2020). Several studies have proposed that promoter reactivation suppression is essential for controlling transcriptional bursting (Zoller *et al.*, 2015), and promoter and enhancer elements regulate burst size and frequency, respectively (Fukaya, Lim and Levine, 2016b) specifically with H3K27ac mark being linearly associated with differential burst frequency (Larsson *et al.*, 2019). In contrast to this, few studies have indicated that enhancers are essential in regulating burst size, both by increasing bursting amplitude, equivalent to the rate of transcription initiation, and bursting length, indicative of the total time the enhancer stays in the ON state (Falo-Sanjuan *et al.*, 2019). However, the molecular mechanism behind cis-regulatory elements (such as the TATA box) and chromatin accessibility at the core promoter in regulating transcriptional bursting kinetics has not been extensively explored yet except for a few studies providing a link between these factors and transcriptional dynamics (Hornung *et al.*, 2012b) (Zoller *et al.*, 2015).

On these lines, my thesis work attempts at shedding light on understanding the potential impact of chromatin remodelling on transcriptional dynamics in context of disease initiation. This work suggests that the cell-to-cell transcriptional variability observed upon loss of *Kat2a* can accelerate pre-leukaemia transformation which further aids in leukaemia progression. Further, this enhanced cell-to-cell transcriptional variability may be associated with a gain in chromatin accessibility. Furthermore, the work highlights that the cell-to-cell transcriptional variability observed during early stages of pre-leukaemia transformation creates a cellular diversity, as evident from differentiation towards B-cell and monocytic lineages, with an increased accessibility towards cell states prone to transformation. These candidate cell populations upon transformation may benefit from low biosynthetic activity that promotes their progression to a leukaemic state (Fig 7.1). This work provides a benchmark in associating chromatin remodelling with transcriptional landscape during pre-leukaemia and has the potential to be further extended to other disease models including disease relapse.

Future work may involve studying the transcriptional bursting parameters upon inhibition/loss of KAT2A during pre-leukaemia progression. This can be achieved by employing techniques like single-molecule RNA FISH (smRNA-FISH) or Droplet digital PCR analysis for KAT2A targets at different time points during *RUNX1-RUNX1T1(9a)* pre-leukaemia progression. As inhibition of KAT2A altered the crosstalk between promoter and candidate enhancer regions, it would be interesting to study how these changes in epigenomic landscape regulate the burst size and burst frequency of *Kat2a* associated targets. Given that loss of *Kat2a* reduced the burst frequency of its associated targets in case of *MLL-AF9* model of leukaemia (Domingues *et al.*, 2020), understanding the transcriptional parameters during *RUNX1-RUNX1T1(9a)* pre-leukaemia would enable a detailed understanding of *Kat2a* mediated transcriptional variability in a disease stage specific manner. A change in burst size may indicate the potential promoter mediated regulation of *Kat2a* targets contributing to transcriptional variability, whereas alterations in burst frequency may suggest crosstalk with putative enhancer regions leading to enhanced transcriptional variability upon *Kat2a* loss.

Although increase in transcriptional instability may seem contrary to an enhanced chromatin accessibility observed upon loss of *Kat2a,* however, this could potentially explain the observations related to cellular differentiation upon *Kat2a* loss as well as initiation of new gene programmes during stochastic speciation of pre-leukaemia cells. An alternative explanation for increase in chromatin accessibility could be a loss of genome integrity by aberrant recruitment of replication licensing factors (Devbhandari *et al.*, 2017; Kurat *et al.*, 2017). With this in mind, future work could explore the applications of targeted DNA sequencing to study signs of DNA damage in candidate regulatory regions associated with *Kat2a* target genes. For this, the upstream regulatory regions of *Kat2a* candidate target genes could be amplified from *RUNX1-RUNX1T1(9a)* pre-leukaemia samples using PCR, and DNA sequencing could be performed on the PCR amplicons. The sequencing data thus generated could be used to study somatic variants accumulated in these regulatory regions during the process of pre-leukaemia progression. An increase in accumulation of somatic variants upon loss of *Kat2a* would suggest an association between loss of *Kat2a* and increase in genomic instability.

Going forward, it will also be interesting to make use of epigenetic tools (e.g. dCas9-KAT2A fusions) being developed in our lab to directly manipulate candidate gene promoters/putative

enhancers and test molecular and functional consequences of the manipulation. To achieve this, guide RNAs against candidate regions of interest could be specifically designed to directly manipulate *Kat2a* mediated acetylation activity at those regions in representative cell lines like Kasumi-1. This will allow study of locus specific changes in transcriptional bursting parameters upon directing *Kat2a* mediated acetylation. This strategy will be instrumental in developing a mechanistic understanding of individual acetylation events and the ability of specific sequences to regulate transcriptional bursting activity. In the future, this knowledge may aid in the development of early diagnostic tools and suggest bespoke therapeutic interventions.

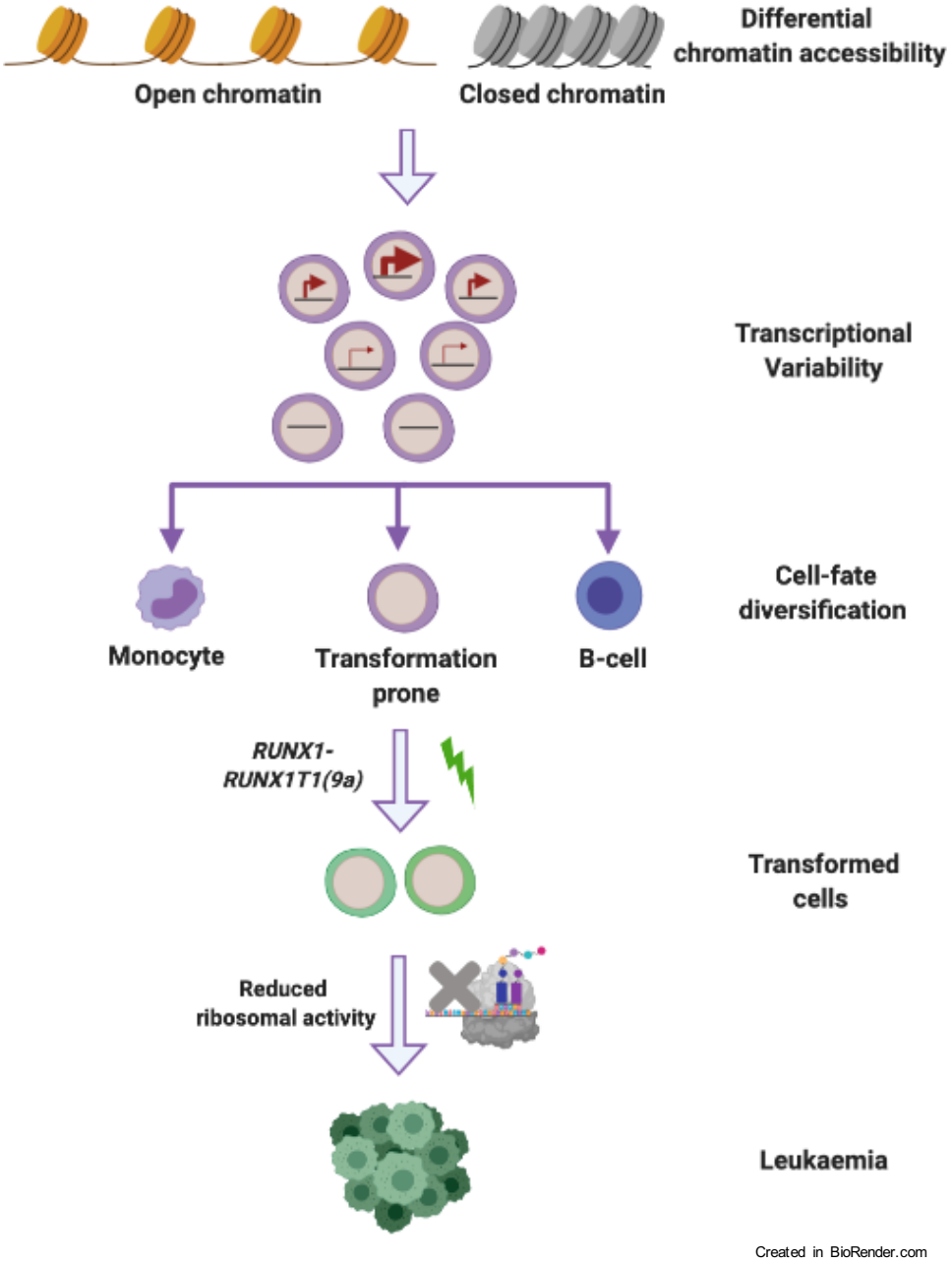

Differential chromatin accessibility
Open chromatin
Closed chromatin
Transcriptional Variability
Cell-fate diversification
Monocyte
Transformation prone
B-cell
RUNX1-RUNX1T1(9a)
Transformed cells
Reduced ribosomal activity
Leukaemia
Created in BioRender.com

Figure 7.1 : Proposed model depicting transcriptional variability consequent to differential chromatin accessibility promoting leukaemia.

The differential chromatin remodelling pattern obtained upon MB-3 treatment of Kasumi-1 cells suggesting an increase in cell-to-cell transcriptional variability (yellow coloured cluster was associated with enhanced accessibility), similar to the observation from scRNA-seq data obtained from *RUNX1-RUNX1T1(9a)* pre-leukaemia. This increase in transcriptional variability leads to cell-fate diversification making transformation prone cells more accessible. These cells upon transformation through *RUNX1-RUNX1T1(9a)* overexpression then benefit though reduced ribosomal activity in order to progress and consequently develop leukaemia.

8 References

Alexandrov, L. B. *et al.* (2013) 'Signatures of mutational processes in human cancer', *Nature*. Nature Publishing Group, 500(7463), pp. 415–421. doi: 10.1038/nature12477.

Alter, B. P. (2003) 'Cancer in Fanconi anemia, 1927-2001', *Cancer*. John Wiley and Sons Inc., 97(2), pp. 425–440. doi: 10.1002/cncr.11046.

Amary, M. F. *et al.* (2011) 'IDH1 and IDH2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours', *Journal of Pathology*. J Pathol, 224(3), pp. 334–343. doi: 10.1002/path.2913.

Amberger, J. S. *et al.* (2015) 'OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders', *Nucleic Acids Research*. Oxford University Press, 43(D1), pp. D789–D798. doi: 10.1093/nar/gku1205.

Anderson, K. *et al.* (2011) 'Genetic variegation of clonal architecture and propagating cells in leukaemia', *Nature*. Nature, 469(7330), pp. 356–361. doi: 10.1038/nature09650.

Angermueller, C. *et al.* (2016) 'Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity', *Nature Methods*. Nature Publishing Group, 13(3), pp. 229–232. doi: 10.1038/nmeth.3728.

Arede, L. *et al.* (2020) 'Unique roles of ATAC and SAGA - KAT2A complexes in normal and malignant hematopoiesis', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2020.05.14.096057. doi: 10.1101/2020.05.14.096057.

Arede, L. and Pina, C. (2020) 'Buffering noise: KAT2A modular contributions to stabilization of transcription and cell identity in cancer and development', *Experimental Hematology*. Elsevier Inc. doi: 10.1016/j.exphem.2020.10.003.

Arezi, B. and Hogrefe, H. (2009) 'Novel mutations in Moloney Murine Leukemia Virus reverse transcriptase increase thermostability through tighter binding to template-primer', *Nucleic Acids Research*. Nucleic Acids Res, 37(2), pp. 473–481. doi: 10.1093/nar/gkn952.

Arias, A. M. and Hayward, P. (2006) 'Filtering transcriptional noise during development:

Concepts and mechanisms', *Nature Reviews Genetics*. Nature Publishing Group, pp. 34–44. doi: 10.1038/nrg1750.

Armour, S. M. *et al.* (2013) 'A High-Confidence Interaction Map Identifies SIRT1 as a Mediator of Acetylation of USP22 and the SAGA Coactivator Complex', *Molecular and Cellular Biology*. American Society for Microbiology, 33(8), pp. 1487–1502. doi: 10.1128/mcb.00971-12.

Artandi, S. E. and DePinho, R. A. (2009) 'Telomeres and telomerase in cancer', *Carcinogenesis*. Oxford University Press, pp. 9–18. doi: 10.1093/carcin/bgp268.

Babushok, D. V. and Bessler, M. (2015) 'Genetic predisposition syndromes: When should they be considered in the work-up of MDS?', *Best Practice and Research: Clinical Haematology*. Bailliere Tindall Ltd, pp. 55–68. doi: 10.1016/j.beha.2014.11.004.

Baca, S. C. *et al.* (2013) 'Punctuated evolution of prostate cancer genomes', *Cell*. Cell, 153(3), pp. 666–677. doi: 10.1016/j.cell.2013.03.021.

Baker, S. C. *et al.* (2005) 'The external RNA controls consortium: A progress report', *Nature Methods*. Nat Methods, 2(10), pp. 731–734. doi: 10.1038/nmeth1005-731.

Baker, S. M. *et al.* (2019) 'Classifying cells with Scasat, a single-cell ATAC-seq analysis tool', *Nucleic Acids Research*. Oxford University Press, 47(2). doi: 10.1093/nar/gky950.

Bakshi, R. *et al.* (2008) 'The leukemogenic t(8;21) fusion protein AML1-ETO controls rRNA genes and associates with nucleolar-organizing regions at mitotic chromosomes', *Journal of Cell Science*. The Company of Biologists Ltd, 121(23), pp. 3981–3990. doi: 10.1242/jcs.033431.

Barkai, N. and Leibler, S. (2000) 'Circadian clocks limited by noise', *Nature*. Macmillan Magazines Ltd, 403(6767), pp. 267–268. doi: 10.1038/35002258.

Basheer, F. *et al.* (2019) 'Contrasting requirements during disease evolution identify EZH2 as a therapeutic target in AML', *Journal of Experimental Medicine*. Rockefeller University Press, 216(4), pp. 966–981. doi: 10.1084/jem.20181276.

Bastide, A. and David, A. (2018) 'The ribosome, (slow) beating heart of cancer (stem) cell',

Oncogenesis. Nature Publishing Group, 7(4). doi: 10.1038/s41389-018-0044-8.

Bäumer, N. *et al.* (2014) 'Maintenance of Leukemia-Initiating Cells Is Regulated by the CDK Inhibitor Inca1', *PLoS ONE*. Edited by Z. Ivanovic. Public Library of Science, 9(12), p. e115578. doi: 10.1371/journal.pone.0115578.

Baylin, S. B. and Jones, P. A. (2011) 'A decade of exploring the cancer epigenome-biological and translational implications', *Nature Reviews Cancer*. Nature Publishing Group, pp. 726–734. doi: 10.1038/nrc3130.

Becskei, A., Kaufmann, B. B. and Van Oudenaarden, A. (2005) 'Contributions of low molecule number and chromosomal positioning to stochastic gene expression', *Nature Genetics*. Nat Genet, 37(9), pp. 937–944. doi: 10.1038/ng1616.

Beerenwinkel, N. *et al.* (2007) 'Genetic progression and the waiting time to cancer', *PLoS Computational Biology*. PLoS Comput Biol, 3(11), pp. 2239–2246. doi: 10.1371/journal.pcbi.0030225.

Beerman, I. *et al.* (2010) 'Stem cells and the aging hematopoietic system', *Current Opinion in Immunology*. Curr Opin Immunol, pp. 500–506. doi: 10.1016/j.coi.2010.06.007.

Bejar, R. *et al.* (2011) 'Clinical Effect of Point Mutations in Myelodysplastic Syndromes', *New England Journal of Medicine*. Massachusetts Medical Society, 364(26), pp. 2496–2506. doi: 10.1056/nejmoa1013343.

Bejar, R. and Steensma, D. P. (2014) 'Recent developments in myelodysplastic syndromes', *Blood*. American Society of Hematology, pp. 2793–2803. doi: 10.1182/blood-2014-04-522136.

Belo, H. *et al.* (2015) 'Epigenetic Alterations in Fanconi anaemia: Role in pathophysiology and therapeutic potential', *PLoS ONE*. Public Library of Science, 10(10), p. 139740. doi: 10.1371/journal.pone.0139740.

Ben-Ami, O. *et al.* (2013) 'Addiction of t(8;21) and inv(16) Acute Myeloid Leukemia to Native RUNX1', *Cell Reports*. Cell Rep, 4(6), pp. 1131–1143. doi: 10.1016/j.celrep.2013.08.020.

Ben-Shem, A. *et al.* (2011) 'The structure of the eukaryotic ribosome at 3.0 Å resolution',

Science. American Association for the Advancement of Science, 334(6062), pp. 1524–1529. doi: 10.1126/science.1212642.

Bengtsson, M. *et al.* (2005) 'Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels', *Genome Research*. Cold Spring Harbor Laboratory Press, 15(10), pp. 1388–1392. doi: 10.1101/gr.3820805.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)*. Wiley, 57(1), pp. 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Berger, M. F. *et al.* (2011) 'The genomic complexity of primary human prostate cancer', *Nature*. Nature, 470(7333), pp. 214–220. doi: 10.1038/nature09744.

Berger, S. L. *et al.* (2009) 'An operational definition of epigenetics', *Genes and Development*. Cold Spring Harbor Laboratory Press, 23(7), pp. 781–783. doi: 10.1101/gad.1787609.

Beyer, K. *et al.* (1998) 'When is "nearest neighbor" meaningful?', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 217–235. doi: 10.1007/3-540-49257-7_15.

Bian, C. *et al.* (2011) 'Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation', *The EMBO Journal*. John Wiley & Sons, Ltd, 30(14), pp. 2829–2842. doi: 10.1038/emboj.2011.193.

Bird, A. (2007) 'Perceptions of epigenetics', *Nature 2007 447:7143*. Nature Publishing Group, 447(7143), pp. 396–398. doi: 10.1038/nature05913.

Blake, W. J. *et al.* (2003) 'Noise in eukaryotic gene expression', *Nature*. Nature, 422(6932), pp. 633–637. doi: 10.1038/nature01546.

Blake, W. J. *et al.* (2006a) 'Phenotypic Consequences of Promoter-Mediated Transcriptional Noise', *Molecular Cell*. Mol Cell, 24(6), pp. 853–865. doi: 10.1016/j.molcel.2006.11.003.

Blake, W. J. *et al.* (2006b) 'Phenotypic Consequences of Promoter-Mediated Transcriptional Noise', *Molecular Cell*. Mol Cell, 24(6), pp. 853–865. doi: 10.1016/j.molcel.2006.11.003.

Block, M., Jacobson, L. O. and Bethard, W. F. (1953) 'Preleukemic acute human leukemia', *Journal of the American Medical Association*. American Medical Association, 152(11), pp. 1018–1028. doi: 10.1001/jama.1953.03690110032010.

Blondel, V. D. (2008) 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics:Theory and Experiment*, 2008.

Boczonadi, V. and Horvath, R. (2014) 'Mitochondria: Impaired mitochondrial translation in human disease', *International Journal of Biochemistry and Cell Biology*. Elsevier Ltd, pp. 77–84. doi: 10.1016/j.biocel.2013.12.011.

Bortoluzzi, S. *et al.* (2002) 'Differential expression of genes coding of ribosomal proteins in different human tissues', *Bioinformatics*. Oxford University Press, 17(12), pp. 1152–1157. doi: 10.1093/bioinformatics/17.12.1152.

Bosc, C., Selak, M. A. and Sarry, J. E. (2017) 'Resistance Is Futile: Targeting Mitochondrial Energetics and Metabolism to Overcome Drug Resistance in Cancer Treatment', *Cell Metabolism*. Cell Press, pp. 705–707. doi: 10.1016/j.cmet.2017.10.013.

Boutros, P. C. *et al.* (2015) 'Spatial genomic heterogeneity within localized, multifocal prostate cancer', *Nature Genetics*. Nature Publishing Group, 47(7), pp. 736–745. doi: 10.1038/ng.3315.

Brady, G., Barbara, M. and Iscove, N. (1990) 'Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies', *Biology*.

Bray, F. *et al.* (2018) 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: A Cancer Journal for Clinicians*. Wiley, 68(6), pp. 394–424. doi: 10.3322/caac.21492.

Bretones, G. *et al.* (2018) 'Altered patterns of global protein synthesis and translational fidelity in RPS15-mutated chronic lymphocytic leukemia', *Blood*. American Society of Hematology, 132(22), pp. 2375–2388. doi: 10.1182/blood-2017-09-804401.

Brownell, J. E. *et al.* (1996) 'Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation', *Cell*. Cell Press, 84(6), pp. 843–851. doi: 10.1016/S0092-8674(00)81063-6.

De Bruin, E. C. *et al.* (2014) ‘Spatial and temporal diversity in genomic instability processes defines lung cancer evolution’, *Science*. American Association for the Advancement of Science, 346(6206), pp. 251–256. doi: 10.1126/science.1253462.

Buenrostro, J. D. *et al.* (2015a) ‘Single-cell chromatin accessibility reveals principles of regulatory variation’, *Nature*. Nature Publishing Group, 523(7561), pp. 486–490. doi: 10.1038/nature14590.

Buenrostro, J. D. *et al.* (2015b) ‘Single-cell chromatin accessibility reveals principles of regulatory variation’, *Nature*. Nature Publishing Group, 523(7561), pp. 486–490. doi: 10.1038/nature14590.

Buganim, Y. *et al.* (2012) ‘Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase’, *Cell*. NIH Public Access, 150(6), pp. 1209–1222. doi: 10.1016/j.cell.2012.08.023.

Busque, L. *et al.* (2012) ‘Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis’, *Nature Genetics*. Nat Genet, 44(11), pp. 1179–1181. doi: 10.1038/ng.2413.

Butler, A. *et al.* (2018) ‘Integrating single-cell transcriptomic data across different conditions, technologies, and species’, *Nature Biotechnology*. Nature Publishing Group, 36(5), pp. 411–420. doi: 10.1038/nbt.4096.

Cahill, D. P. *et al.* (1999) ‘Genetic instability and darwinian selection in tumours’, *Trends in Cell Biology*. Elsevier Current Trends, pp. M57–M60. doi: 10.1016/S0962-8924(99)01661-X.

Cai, L., Friedman, N. and Xie, X. S. (2006) ‘Stochastic protein expression in individual cells at the single molecule level’, *Nature*. Nature Publishing Group, 440(7082), pp. 358–362. doi: 10.1038/nature04599.

Cai, X., Gao, L., Teng, L., Ge, J., *et al.* (2015) ‘Runx1 Deficiency Decreases Ribosome Biogenesis and Confers Stress Resistance to Hematopoietic Stem and Progenitor Cells’, *Cell Stem Cell*. Cell Press, 17(2), pp. 165–177. doi: 10.1016/j.stem.2015.06.002.

Cai, X., Gao, L., Teng, L., Mason, P. J., *et al.* (2015) ‘Runx1 Deficiency Decreases Ribosome

Biogenesis and Confers Stress Resistance to Hematopoietic Stem and Progenitor Cells', *Stem Cell*, 17, pp. 165–177. doi: 10.1016/j.stem.2015.06.002.

Campbell, P. J. *et al.* (2010) 'The patterns and dynamics of genomic instability in metastatic pancreatic cancer', *Nature*. NIH Public Access, 467(7319), pp. 1109–1113. doi: 10.1038/nature09460.

Carroll, C. J. *et al.* (2013) 'Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mitochondrial infantile cardiomyopathy', *Journal of Medical Genetics*. J Med Genet, 50(3), pp. 151–159. doi: 10.1136/jmedgenet-2012-101375.

Catlin, S. N. *et al.* (2011) 'The replication rate of human hematopoietic stem cells in vivo', *Blood*. Blood, 117(17), pp. 4460–4466. doi: 10.1182/blood-2010-08-303537.

Cattell RB (1966) 'The Scree Test For The Number Of Factors', *Multivariate Behav Res*, 1(2), pp. 245–76. doi: 10.1207/s15327906mbr0102_10.

Chalancon, G. *et al.* (2012) 'Interplay between gene expression noise and regulatory network architecture', *Trends in Genetics*. Elsevier Current Trends, pp. 221–232. doi: 10.1016/j.tig.2012.01.006.

Chan, K. L., North, P. S. and Hickson, I. D. (2007) 'BLM is required for faithful chromosome segregation and its localization defines a class of ultrafine anaphase bridges', *EMBO Journal*. European Molecular Biology Organization, 26(14), pp. 3397–3409. doi: 10.1038/sj.emboj.7601777.

Chan, W.-I. *et al.* (2011) 'The Transcriptional Coactivator Cbp Regulates Self-Renewal and Differentiation in Adult Hematopoietic Stem Cells', *Molecular and Cellular Biology*. American Society for Microbiology, 31(24), pp. 5046–5060. doi: 10.1128/mcb.05830-11.

Chang, H. H. *et al.* (2008) 'Transcriptome-wide noise controls lineage choice in mammalian progenitor cells', *Nature*. Nature Publishing Group, 453(7194), pp. 544–547. doi: 10.1038/nature06965.

Chen, J. *et al.* (2010) 'Pygo2 Associates with MLL2 Histone Methyltransferase and GCN5 Histone Acetyltransferase Complexes To Augment Wnt Target Gene Expression and Breast

Cancer Stem-Like Cell Expansion', *Molecular and Cellular Biology*. American Society for Microbiology, 30(24), pp. 5621–5635. doi: 10.1128/mcb.00465-10.

Chen, L. *et al.* (2013) 'Lysine acetyltransferase GCN5 potentiates the growth of non-small cell lung cancer via promotion of E2F1, cyclin D1, and cyclin E1 expression', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 288(20), pp. 14510–14521. doi: 10.1074/jbc.M113.458737.

Chen, W. L. *et al.* (2014) 'A distinct glucose metabolism signature of acute myeloid leukemia with prognostic value', *Blood*. American Society of Hematology, 124(10), pp. 1645–1654. doi: 10.1182/blood-2014-02-554204.

Choesmel, V. *et al.* (2007) 'Impaired ribosome biogenesis in Diamond-Blackfan anemia', *Blood*. Blood, 109(3), pp. 1275–1283. doi: 10.1182/blood-2006-07-038372.

Chou, W. C. *et al.* (2010) 'Distinct clinical and biologic characteristics in adult acute myeloid leukemia bearing the isocitrate dehydrogenase 1 mutation', *Blood*. Blood, 115(14), pp. 2749–2754. doi: 10.1182/blood-2009-11-253070.

Churpek, J. E. *et al.* (2015) 'Genomic analysis of germ line and somatic variants in familial myelodysplasia/acute myeloid leukemia', *Blood*. American Society of Hematology, 126(22), pp. 2484–2490. doi: 10.1182/blood-2015-04-641100.

Clark, S. J. and Melki, J. (2002) 'DNA methylation and gene silencing in cancer: Which is the guilty party?', *Oncogene*. Oncogene, pp. 5380–5387. doi: 10.1038/sj.onc.1205598.

Corces-Zimmerman, M. R. *et al.* (2014a) 'Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(7), pp. 2548–2553. doi: 10.1073/pnas.1324297111.

Corces-Zimmerman, M. R. *et al.* (2014b) 'Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(7), pp. 2548–2553. doi: 10.1073/pnas.1324297111.

Corces-Zimmerman, M. R. *et al.* (2014c) 'Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(7), pp. 2548–2553. doi: 10.1073/pnas.1324297111.

Corral, J. *et al.* (1996) 'An MII-AF9 fusion gene made by homologous recombination causes acute leukemia in chimeric mice: A method to create fusion oncogenes', *Cell*. Cell Press, 85(6), pp. 853–861. doi: 10.1016/S0092-8674(00)81269-6.

Corrigan, A. M. *et al.* (2016) 'A continuum model of transcriptional bursting', *eLife*. eLife Sciences Publications Ltd, 5(FEBRUARY2016). doi: 10.7554/eLife.13051.

Crasta, K. *et al.* (2012) 'DNA breaks and chromosome pulverization from errors in mitosis', *Nature*. Nature Publishing Group, pp. 53–58. doi: 10.1038/nature10802.

Cunningham, I. and Kohno, B. (2016) '18FDG-PET/CT: 21st century approach to leukemic tumors in 124 cases', *American Journal of Hematology*. Wiley-Liss Inc., 91(4), pp. 379–384. doi: 10.1002/ajh.24287.

Dang, L. *et al.* (2009) 'Cancer-associated IDH1 mutations produce 2-hydroxyglutarate', *Nature*. Nature Publishing Group, 462(7274), pp. 739–744. doi: 10.1038/nature08617.

Dar, R. D. *et al.* (2012) 'Transcriptional burst frequency and burst size are equally modulated across the human genome', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 109(43), pp. 17454–17459. doi: 10.1073/pnas.1213530109.

Davis, A., Gao, R. and Navin, N. (2017) 'Tumor evolution: Linear, branching, neutral or punctuated?', *Biochimica et Biophysica Acta - Reviews on Cancer*. Elsevier B.V., 1867(2), pp. 151–161. doi: 10.1016/j.bbcan.2017.01.003.

Delmans, M. and Hemberg, M. (2016) 'Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data', *BMC Bioinformatics*. BioMed Central Ltd., 17(1), p. 110. doi: 10.1186/s12859-016-0944-6.

Devbhandari, S. *et al.* (2017) 'Chromatin Constrains the Initiation and Elongation of DNA

Replication', *Molecular Cell*. Cell Press, 65(1), pp. 131–141. doi: 10.1016/j.molcel.2016.10.035.

Devlin, E. E. *et al.* (2010) 'A transgenic mouse model demonstrates a dominant negative effect of a point mutation in the RPS19 gene associated with Diamond-Blackfan anemia', *Blood*. American Society of Hematology, 116(15), pp. 2826–2835. doi: 10.1182/blood-2010-03-275776.

Ding, L. *et al.* (2012) 'Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing', *Nature*. Nature Publishing Group, 481(7382), pp. 506–510. doi: 10.1038/nature10738.

Dixon, J. R., Gorkin, D. U. and Ren, B. (2016) 'Chromatin Domains: The Unit of Chromosome Organization', *Molecular Cell*. Cell Press, pp. 668–680. doi: 10.1016/j.molcel.2016.05.018.

Djabali, M. *et al.* (1992) 'A trithorax–like gene is interrupted by chromosome 11q23 translocations in acute leukaemias', *Nature Genetics*. Nat Genet, 2(2), pp. 113–118. doi: 10.1038/ng1092-113.

Domingues, A. F. *et al.* (2020) 'Loss of KAT2A enhances transcriptional noise and depletes acute myeloid leukemia stem-like cells', *eLife*. eLife Sciences Publications Ltd, 9. doi: 10.7554/eLife.51754.

Doulatov, S. *et al.* (2012) 'Hematopoiesis: A human perspective', *Cell Stem Cell*. Cell Press, pp. 120–136. doi: 10.1016/j.stem.2012.01.006.

Duttke, S. H. *et al.* (2019) 'Identification and dynamic quantification of regulatory elements using total RNA'. doi: 10.1101/gr.253492.119.

Eberwine, J. *et al.* (1992) 'Analysis of gene expression in single live neurons', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 89(7), pp. 3010–3014. doi: 10.1073/pnas.89.7.3010.

Eldar, A. and Elowitz, M. B. (2010) 'Functional roles for noise in genetic circuits', *Nature*. Nature Publishing Group, pp. 167–173. doi: 10.1038/nature09326.

Elowitz, M. B. *et al.* (2002) 'Stochastic gene expression in a single cell', *Science*. American

Association for the Advancement of Science, 297(5584), pp. 1183–1186. doi: 10.1126/science.1070919.

Enver, T., Heyworth, C. M. and Dexter, T. M. (1998) 'Do stem cells play dice?', *Blood*. doi: 10.1182/blood.v92.2.348.con3_348_351.

Eriksson, A., Lennartsson, A. and Lehmann, S. (2015) 'Epigenetic aberrations in acute myeloid leukemia: Early key events during leukemogenesis', *Experimental Hematology*. Elsevier Inc., pp. 609–624. doi: 10.1016/j.exphem.2015.05.009.

Ewing, B. and Green, P. (1998) 'Base-calling of automated sequencer traces using phred. II. Error probabilities', *Genome Research*. Cold Spring Harbor Laboratory Press, 8(3), pp. 186–194. doi: 10.1101/gr.8.3.186.

Falo-Sanjuan, J. *et al.* (2019) 'Enhancer Priming Enables Fast and Sustained Transcriptional Responses to Notch Signaling', *Developmental Cell*. Cell Press, 50(4), pp. 411-425.e8. doi: 10.1016/j.devcel.2019.07.002.

Farge, T. *et al.* (2017) 'Chemotherapy-resistant human acute myeloid leukemia cells are not enriched for leukemic stem cells but require oxidative metabolism', *Cancer Discovery*. American Association for Cancer Research Inc., 7(7), pp. 716–735. doi: 10.1158/2159-8290.CD-16-0441.

Farrar, J. E. *et al.* (2014) 'Exploiting pre-rRNA processing in Diamond Blackfan anemia gene discovery and diagnosis', *American Journal of Hematology*. Wiley-Liss Inc., 89(10), pp. 985–991. doi: 10.1002/ajh.23807.

Fathi, A. T. *et al.* (2015) 'Biochemical, epigenetic, and metabolic approaches to target IDH mutations in acute myeloid leukemia', *Seminars in Hematology*. W.B. Saunders, pp. 165–171. doi: 10.1053/j.seminhematol.2015.03.002.

Fearon, E. R. and Vogelstein, B. (1990) 'A genetic model for colorectal tumorigenesis', *Cell*. Cell, pp. 759–767. doi: 10.1016/0092-8674(90)90186-I.

Feinstein, P. G. *et al.* (1995) *Identification of Homeotic Target Genes i n Drosophila melanogastm Including neruy, a Proto-Oncogene Homologue*.

Feldser, D. M. *et al.* (2010) 'Stage-specific sensitivity to p53 restoration during lung cancer progression', *Nature*, 468(7323), pp. 572–575. doi: 10.1038/nature09535.

Feron, O. (2009) 'Pyruvate into lactate and back: From the Warburg effect to symbiotic energy fuel exchange in cancer cells', *Radiotherapy and Oncology*. Radiother Oncol, pp. 329–333. doi: 10.1016/j.radonc.2009.06.025.

Fialkow, P. J., Janssen, J. W. G. and Bartram, C. R. (1991) 'Clonal remissions in acute nonlymphocytic leukemia: Evidence for a multistep pathogenesis of the malignany', *Blood*. Blood, 77(7), pp. 1415–1417. doi: 10.1182/blood.v77.7.1415.bloodjournal7771415.

Figueroa, M. E. *et al.* (2010) 'Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation', *Cancer Cell*. Cancer Cell, 18(6), pp. 553–567. doi: 10.1016/j.ccr.2010.11.015.

Flygare, J. *et al.* (2007) 'Human RPS19, the gene mutated in Diamond-Blackfan anemia, encodes a ribosomal protein required for the maturation of 40S ribosomal subunits', *Blood*. Blood, 109(3), pp. 980–986. doi: 10.1182/blood-2006-07-038232.

Ford, A. M. *et al.* (1993) 'In utero rearrangements in the trithorax-related oncogene in infant leukaemias', *Nature*. Nature, 363(6427), pp. 358–360. doi: 10.1038/363358a0.

Ford, A. M. *et al.* (1998) 'Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 95(8), pp. 4584–4588. doi: 10.1073/pnas.95.8.4584.

Forster, V. J. *et al.* (2016) 'The leukemia-associated RUNX1/ETO oncoprotein confers a mutator phenotype', *Leukemia*. Nature Publishing Group, pp. 250–253. doi: 10.1038/leu.2015.133.

Fraga, M. F. *et al.* (2005) 'Epigenetic differences arise during the lifetime of monozygotic twins', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 102(30), pp. 10604–10609. doi: 10.1073/pnas.0500398102.

Frieda, K. L. *et al.* (2017) 'Synthetic recording and in situ readout of lineage information in single cells', *Nature*. Nature Publishing Group, 541(7635), pp. 107–111. doi:

10.1038/nature20777.

Fudenberg, G. *et al.* (2016) 'Formation of Chromosomal Domains by Loop Extrusion', *Cell Reports*. Elsevier B.V., 15(9), pp. 2038–2049. doi: 10.1016/j.celrep.2016.04.085.

Fukaya, T., Lim, B. and Levine, M. (2016a) 'Enhancer Control of Transcriptional Bursting', *Cell*. Cell Press, 166(2), pp. 358–368. doi: 10.1016/j.cell.2016.05.025.

Fukaya, T., Lim, B. and Levine, M. (2016b) 'Enhancer Control of Transcriptional Bursting', *Cell*. Cell Press, 166(2), pp. 358–368. doi: 10.1016/j.cell.2016.05.025.

Fukuda, R. *et al.* (2007) 'HIF-1 Regulates Cytochrome Oxidase Subunits to Optimize Efficiency of Respiration in Hypoxic Cells', *Cell*. Cell, 129(1), pp. 111–122. doi: 10.1016/j.cell.2007.01.047.

Fukuyama, T. *et al.* (2001) 'MTG8 proto-oncoprotein interacts with the regulatory subunit of type II cyclic AMP-dependent protein kinase in lymphocytes', *Oncogene*. Oncogene, 20(43), pp. 6225–6232. doi: 10.1038/sj.onc.1204794.

Gale, R. E. *et al.* (1993) 'Frequency of clonal remission in acute myeloid leukaemia', *The Lancet*. Lancet, 341(8838), pp. 138–142. doi: 10.1016/0140-6736(93)90004-Z.

Galmiche, L. *et al.* (2011) 'Exome sequencing identifies MRPL3 mutation in mitochondrial cardiomyopathy', *Human Mutation*. Hum Mutat, 32(11), pp. 1225–1231. doi: 10.1002/humu.21562.

Ganem, N. J., Storchova, Z. and Pellman, D. (2007) 'Tetraploidy, aneuploidy and cancer', *Current Opinion in Genetics and Development*. Curr Opin Genet Dev, pp. 157–162. doi: 10.1016/j.gde.2007.02.011.

Gao, J. *et al.* (1991) 'Isolation of a yeast artificial chromosome spanning the 8;21 translocation breakpoint t(8;21)(q22;q22.3) in acute myelogenous leukemia', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 88(11), pp. 4882–4886. doi: 10.1073/pnas.88.11.4882.

Garg, M. *et al.* (2015) 'Profiling of somatic mutations in acute myeloid leukemia with FLT3-ITD at diagnosis and relapse', *Blood*. American Society of Hematology, 126(22), pp. 2491–

2501. doi: 10.1182/blood-2015-05-646240.

Gaur, R. *et al.* (2008) ‘A Single Mammalian Mitochondrial Translation Initiation Factor Functionally Replaces Two Bacterial Factors’, *Molecular Cell*. NIH Public Access, 29(2), pp. 180–190. doi: 10.1016/j.molcel.2007.11.021.

Gawad, C., Koh, W. and Quake, S. R. (2014) ‘Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(50), pp. 17947–17952. doi: 10.1073/pnas.1420822111.

Genovese, G. *et al.* (2014) ‘Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence’, *New England Journal of Medicine*. Massachusetts Medical Society, 371(26), pp. 2477–2487. doi: 10.1056/nejmoa1409405.

Gerlinger, Marco,Nicholas Mcgranahan, C. S. (2014) ‘Cancer:Evolution Within a Lifetime’, *The Annual Reviews of Genetics*, 48, pp. 215–36. doi: 10.1146/annurev-genet-120213-092314.

Gerlinger, M. *et al.* (2012a) ‘Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing’, *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 366(10), pp. 883–892. doi: 10.1056/nejmoa1113205.

Gerlinger, M. *et al.* (2012b) ‘Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing’, *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 366(10), pp. 883–892. doi: 10.1056/nejmoa1113205.

Gerlinger, M. and Swanton, C. (2010) ‘How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine’, *British Journal of Cancer*. Nature Publishing Group, pp. 1139–1143. doi: 10.1038/sj.bjc.6605912.

Gomez-Roman, N. *et al.* (2003) ‘Direct activation of RNA polymerase III transcription by c-Myc’, *Nature*. Nature, 421(6920), pp. 290–294. doi: 10.1038/nature01327.

Gordon, D. J., Resio, B. and Pellman, D. (2012) ‘Causes and consequences of aneuploidy in cancer’, *Nature Reviews Genetics*. Nat Rev Genet, pp. 189–203. doi: 10.1038/nrg3123.

Gorgoulis, V. G. *et al.* (2005) ‘Activation of the DNA damage checkpoint and genomic

instability in human precancerous lesions', *Nature*. Nature, 434(7035), pp. 907–913. doi: 10.1038/nature03485.

Gottlieb, E., Vander Heiden, M. G. and Thompson, C. B. (2000) 'Bcl-xL Prevents the Initial Decrease in Mitochondrial Membrane Potential and Subsequent Reactive Oxygen Species Production during Tumor Necrosis Factor Alpha-Induced Apoptosis', *Molecular and Cellular Biology*. American Society for Microbiology, 20(15), pp. 5680–5689. doi: 10.1128/mcb.20.15.5680-5689.2000.

Gould, S. J. and Eldredge, N. (1993) 'Punctuated equilibrium comes of age', *Nature*. Nature, pp. 223–227. doi: 10.1038/366223a0.

Greaves, M. and Maley, C. C. (2012) 'Clonal evolution in cancer', *Nature*. NIH Public Access, pp. 306–313. doi: 10.1038/nature10762.

Greber, B. J. and Ban, N. (2016) 'Structure and Function of the Mitochondrial Ribosome'. doi: 10.1146/annurev-biochem-060815-014343.

Grebien, F. *et al.* (2015) 'Pharmacological targeting of the Wdr5-MLL interaction in C/EBPα N-terminal leukemia', *Nature Chemical Biology*. Nature Publishing Group, 11(8), pp. 571–578. doi: 10.1038/nchembio.1859.

Green, A. S. *et al.* (2010) 'The LKB1/AMPK signaling pathway has tumor suppressor activity in acute myeloid leukemia through the repression of mTOR-dependent oncogenic mRNA translation', *Blood*. Blood, 116(20), pp. 4262–4273. doi: 10.1182/blood-2010-02-269837.

Grinev, V. V. *et al.* (2021) 'RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia', *Nature Communications*. Nature Research, 12(1), pp. 1–16. doi: 10.1038/s41467-020-20848-z.

Grove, C. S. and Vassiliou, G. S. (2014) 'Acute myeloid leukaemia: A paradigm for the clonal evolution of cancer?', *DMM Disease Models and Mechanisms*. Company of Biologists Ltd, pp. 941–951. doi: 10.1242/dmm.015974.

Guelman, S. *et al.* (2006) 'Host Cell Factor and an Uncharacterized SANT Domain Protein Are Stable Components of ATAC, a Novel dAda2A/dGcn5-Containing Histone

Acetyltransferase Complex in Drosophila', *Molecular and Cellular Biology*. American Society for Microbiology, 26(3), pp. 871–882. doi: 10.1128/mcb.26.3.871-882.2006.

Guelman, S. *et al.* (2009) 'The Double-Histone-Acetyltransferase Complex ATAC Is Essential for Mammalian Development', *Molecular and Cellular Biology*. American Society for Microbiology, 29(5), pp. 1176–1188. doi: 10.1128/mcb.01599-08.

Hafemeister, C. and Satija, R. (2019) 'Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression', *Genome Biology*. BioMed Central Ltd., 20(1), p. 296. doi: 10.1186/s13059-019-1874-1.

Haferlach, T. *et al.* (2014) 'Landscape of genetic lesions in 944 patients with myelodysplastic syndromes', *Leukemia*. Leukemia, 28(2), pp. 241–247. doi: 10.1038/leu.2013.336.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: The next generation', *Cell*. Cell, pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Harbst, K. *et al.* (2016) 'Multiregion whole-exome sequencing uncovers the genetic evolution and mutational heterogeneity of early-stage metastatic melanoma', *Cancer Research*. American Association for Cancer Research Inc., 76(16), pp. 4765–4774. doi: 10.1158/0008-5472.CAN-15-3476.

Hashimshony, T. *et al.* (2012) 'CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification', *Cell Reports*. Cell Rep, 2(3), pp. 666–673. doi: 10.1016/j.celrep.2012.08.003.

Hatlen, M. A. *et al.* (2016) 'Integrative genetic analysis of mouse and human AML identifies cooperating disease alleles', *Journal of Experimental Medicine*. Rockefeller University Press, 213(1), pp. 25–34. doi: 10.1084/jem.20150524.

Hay, N. (2016a) 'Reprogramming glucose metabolism in cancer: Can it be exploited for cancer therapy?', *Nature Reviews Cancer*. Nature Publishing Group, pp. 635–649. doi: 10.1038/nrc.2016.77.

Hay, N. (2016b) 'Reprogramming glucose metabolism in cancer: Can it be exploited for cancer therapy?', *Nature Reviews Cancer*. Nature Publishing Group, pp. 635–649. doi: 10.1038/nrc.2016.77.

Hayashi, T. *et al.* (2010) 'Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research', *Development, Growth & Differentiation*. John Wiley & Sons, Ltd, 52(1), pp. 131–144. doi: 10.1111/j.1440-169X.2009.01157.x.

Haynes, W. (2013) 'Bonferroni Correction', in *Encyclopedia of Systems Biology*. Springer New York, pp. 154–154. doi: 10.1007/978-1-4419-9863-7_1213.

Hebbes, T. R., Thorne, A. W. and Crane-Robinson, C. (1988) 'A direct link between core histone acetylation and transcriptionally active chromatin.', *The EMBO Journal*. John Wiley & Sons, Ltd, 7(5), pp. 1395–1402. doi: 10.1002/j.1460-2075.1988.tb02956.x.

Hedlund, E. and Deng, Q. (2017) 'Single-cell RNA sequencing: Technical advancements and biological applications'. doi: 10.1016/j.mam.2017.07.003.

Hedlund, E. and Deng, Q. (2018) 'Single-cell RNA sequencing: Technical advancements and biological applications', *Molecular Aspects of Medicine*. Elsevier Ltd, pp. 36–46. doi: 10.1016/j.mam.2017.07.003.

Hemerly, J. P., Bastos, A. U. and Cerutti, J. M. (2010) 'Identification of several novel non-p.R132 IDH1 variants in thyroid carcinomas', *European Journal of Endocrinology*. Eur J Endocrinol, 163(5), pp. 747–755. doi: 10.1530/EJE-10-0473.

Henras, A. K. *et al.* (2008) 'The post-transcriptional steps of eukaryotic ribosome biogenesis', *Cellular and Molecular Life Sciences*. Cell Mol Life Sci, pp. 2334–2359. doi: 10.1007/s00018-008-8027-0.

Herst, P. M. *et al.* (2011) 'The level of glycolytic metabolism in acute myeloid leukemia blasts at diagnosis is prognostic for clinical outcome', *Journal of Leukocyte Biology*. Wiley, 89(1), pp. 51–55. doi: 10.1189/jlb.0710417.

Hidalgo San Jose, L. *et al.* (2020) 'Modest Declines in Proteome Quality Impair Hematopoietic Stem Cell Self-Renewal', *Cell Reports*. Elsevier B.V., 30(1), pp. 69-80.e6. doi: 10.1016/j.celrep.2019.12.003.

Hidalgo San Jose, L. and Signer, R. A. J. (2019) 'Cell-type-specific quantification of protein

synthesis in vivo', *Nature Protocols*. Nature Publishing Group, 14(2), pp. 441–460. doi: 10.1038/s41596-018-0100-z.

Hindson, B. J. *et al.* (2011) 'High-throughput droplet digital PCR system for absolute quantitation of DNA copy number', *Analytical Chemistry*. UTC, 83(22), pp. 8604–8610. doi: 10.1021/ac202028g.

Hitchins, M. P. *et al.* (2007) ' Inheritance of a Cancer-Associated MLH1 Germ-Line Epimutation ', *New England Journal of Medicine*. Massachusetts Medical Society, 356(7), pp. 697–705. doi: 10.1056/nejmoa064522.

Holliday, R. (1987) 'The inheritance of epigenetic defects', *Science*. Science, 238(4824), pp. 163–170. doi: 10.1126/science.3310230.

Hong, S. H. *et al.* (2020) 'Epigenetic Approaches to the Treatment of Renal Cell Cancer', *The Korean Journal of Urological Oncology*. Korean Urological Oncology Society, 18(2), pp. 78–90. doi: 10.22465/kjuo.2020.18.2.78.

Hornung, G. *et al.* (2012a) 'Noise-mean relationship in mutated promoters', *Genome Research*. Genome Res, 22(12), pp. 2409–2417. doi: 10.1101/gr.139378.112.

Hornung, G. *et al.* (2012b) 'Noise-mean relationship in mutated promoters', *Genome Research*. Cold Spring Harbor Laboratory Press, 22(12), pp. 2409–2417. doi: 10.1101/gr.139378.112.

Horos, R. *et al.* (2012) 'Ribosomal deficiencies in Diamond-Blackfan anemia impair translation of transcripts essential for differentiation of murine and human erythroblasts', *Blood*. Blood, 119(1), pp. 262–272. doi: 10.1182/blood-2011-06-358200.

Horton, S. J. *et al.* (2013) 'MLL-AF9-mediated immortalization of human hematopoietic cells along different lineages changes during ontogeny', *Leukemia*. Leukemia, 27(5), pp. 1116–1126. doi: 10.1038/leu.2012.343.

Horton, S. J. (2017) 'Early loss of Crebbp confers malignant stem cell properties on lymphoid progenitors', *Nature Cell Biology*, 19(9), pp. 1093–1104. doi: 10.1038/ncb3597.

Hou, Y. *et al.* (2016) 'Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas', *Cell Research*. Nature Publishing

Group, 26(3), pp. 304–319. doi: 10.1038/cr.2016.23.

Huang, S. (2009) ‘Reprogramming cell fates: Reconciling rarity with robustness’, *BioEssays*. Bioessays, 31(5), pp. 546–560. doi: 10.1002/bies.200800189.

Huntly, B. J. P. *et al.* (2004) ‘MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors’, *Cancer Cell*. Cancer Cell, 6(6), pp. 587–596. doi: 10.1016/j.ccr.2004.10.015.

Huret, J. L. *et al.* (2000) ‘The “Atlas of Genetics and Cytogenetics in Oncology and Haematology” on the internet and a review on infant leukemias’, *Cancer Genetics and Cytogenetics*. Cancer Genet Cytogenet, 120(2), pp. 155–159. doi: 10.1016/S0165-4608(99)00250-2.

Hwang, B., Lee, J. H. and Bang, D. (2018) ‘Single-cell RNA sequencing technologies and bioinformatics pipelines’, *Experimental and Molecular Medicine*. Nature Publishing Group. doi: 10.1038/s12276-018-0071-8.

Isa, A. *et al.* (2018) ‘Identification of glycolytic pathway as RUNX1/ETO-dependent for propagation and survival’, in *31. Jahrestagung der Kind-Philipp-Stiftung für pädiatrisch onkologische Forschung*. Georg Thieme Verlag KG, p. 3. doi: 10.1055/s-0038-1644984.

Islam, S. *et al.* (2011) ‘Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq’, *Genome Research*. Genome Res, 21(7), pp. 1160–1167. doi: 10.1101/gr.110882.110.

Islam, S. *et al.* (2014a) ‘Quantitative single-cell RNA-seq with unique molecular identifiers’, *Nature Methods*. Nat Methods, 11(2), pp. 163–166. doi: 10.1038/nmeth.2772.

Islam, S. *et al.* (2014b) ‘Quantitative single-cell RNA-seq with unique molecular identifiers’, *Nature Methods*. Nat Methods, 11(2), pp. 163–166. doi: 10.1038/nmeth.2772.

Itzykson, R. and Fenaux, P. (2013) ‘Epigenetics of myelodysplastic syndromes’, 28. doi: 10.1038/leu.2013.343.

Jaccard, P. (1901) ‘Distribution de la flore alpine dans le bassin des dranses et dans quelques rgions voisines’, *Bulletin de la Socit Vaudoise des Sciences Naturelles*, (37), pp. 241–272.

Jacobs, A. (1985) ‘Myelodysplastic syndromes: Pathogenesis, functional abnormalities, and clinical implications’, *Journal of Clinical Pathology*. BMJ Publishing Group, pp. 1201–1217. doi: 10.1136/jcp.38.11.1201.

Janssen, A. *et al.* (2011) ‘Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations’, *Science*. American Association for the Advancement of Science, 333(6051), pp. 1895–1898. doi: 10.1126/science.1210214.

Järås, M. and Ebert, B. L. (2011) ‘Power Cut: Inhibiting Mitochondrial Translation to Target Leukemia’, *Cancer Cell*. Cancer Cell, pp. 555–556. doi: 10.1016/j.ccr.2011.10.028.

Ji, Z. *et al.* (2020) ‘Single-cell ATAC-seq signal extraction and enhancement with SCATE’, *Genome Biology*. BioMed Central, 21(1), p. 161. doi: 10.1186/s13059-020-02075-3.

Ji, Z. and Ji, H. (2019) ‘Pseudotime reconstruction using TSCAN’, in *Methods in Molecular Biology*. Humana Press Inc., pp. 115–124. doi: 10.1007/978-1-4939-9057-3_8.

Jiao, B. *et al.* (2009) ‘AML1-ETO9a is correlated with C-KIT overexpression/mutations and indicates poor disease outcome in t(8;21) acute myeloid leukemia-M2’, *Leukemia*. Nature Publishing Group, 23(9), pp. 1598–1604. doi: 10.1038/leu.2009.104.

Jin, Q. *et al.* (2014) ‘Gcn5 and PCAF Regulate PPAR and Prdm16 Expression To Facilitate Brown Adipogenesis’, *Molecular and Cellular Biology*. American Society for Microbiology, 34(19), pp. 3746–3753. doi: 10.1128/mcb.00622-14.

Jo, S. H. *et al.* (2001) ‘Control of Mitochondrial Redox Balance and Cellular Defense against Oxidative Damage by Mitochondrial NADP+-dependent Isocitrate Dehydrogenase’, *Journal of Biological Chemistry*. J Biol Chem, 276(19), pp. 16168–16176. doi: 10.1074/jbc.M010120200.

Johnson, W. E., Li, C. and Rabinovic, A. (2007) ‘Adjusting batch effects in microarray expression data using empirical Bayes methods’, *Biostatistics*. Oxford Academic, 8(1), pp. 118–127. doi: 10.1093/biostatistics/kxj037.

Jones, P. A. (1999) ‘The DNA methylation paradox’, *Trends in Genetics*. Trends Genet, pp. 34–37. doi: 10.1016/S0168-9525(98)01636-9.

Julius, M. H., Masuda, T. and Herzenberg, L. A. (1972) ‘Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter.’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 69(7), pp. 1934–1938. doi: 10.1073/pnas.69.7.1934.

Junttila, M. R. and De Sauvage, F. J. (2013) ‘Influence of tumour micro-environment heterogeneity on therapeutic response’, *Nature*. Nature Publishing Group, pp. 346–354. doi: 10.1038/nature12626.

Kakutani, T. *et al.* (1996) ‘Developmental abnormalities and epimutations associated with DNA hypomethylation mutations’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 93(22), pp. 12406–12411. doi: 10.1073/pnas.93.22.12406.

Kamminga, L. M. *et al.* (2000) ‘Autonomous behavior of hematopoietic stem cells’, *Experimental Hematology*. Exp Hematol, 28(12), pp. 1451–1459. doi: 10.1016/S0301-472X(00)00543-9.

Kampen, K. R. *et al.* (2020) ‘Hallmarks of ribosomopathies’, *Nucleic acids research*. NLM (Medline), pp. 1013–1028. doi: 10.1093/nar/gkz637.

Kasper, L. H. *et al.* (2002) ‘A transcription-factor-binding surface of coactivator p300 is required for haematopoiesis’, *Nature*. Nature, 419(6908), pp. 738–743. doi: 10.1038/nature01062.

Kattih, B. *et al.* (2020) ‘IDH1/2 mutations in acute myeloid leukemia patients and risk of coronary artery disease and cardiac dysfunction—a retrospective propensity score analysis’, *Leukemia*. Springer Nature. doi: 10.1038/s41375-020-01043-x.

De Keersmaecker, K. *et al.* (2013a) ‘Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia’, *Nature Genetics*. Nat Genet, 45(2), pp. 186–190. doi: 10.1038/ng.2508.

De Keersmaecker, K. *et al.* (2013b) ‘Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia’, *Nature Genetics*. Nat Genet, 45(2), pp. 186–190. doi: 10.1038/ng.2508.

De Keersmaecker, K., Sulima, S. O. and Dinman, J. D. (2015) 'Ribosomopathies and the paradox of cellular hypo- to hyperproliferation', *Blood*. American Society of Hematology, pp. 1377–1382. doi: 10.1182/blood-2014-10-569616.

Keller, G. (2005) 'Embryonic stem cell differentiation: Emergence of a new era in biology and medicine', *Genes and Development*. Genes Dev, pp. 1129–1155. doi: 10.1101/gad.1303605.

Kelly, L., Clark, J. and Gilliland, D. G. (2002) 'Comprehensive genotypic analysis of leukemia: Clinical and therapeutic implications', *Current Opinion in Oncology*. Curr Opin Oncol, pp. 10–18. doi: 10.1097/00001622-200201000-00003.

Kern, S. E. (2012) 'Why your new cancer biomarker may never work: Recurrent patterns and remarkable diversity in biomarker failures', *Cancer Research*. Cancer Res, pp. 6097–6101. doi: 10.1158/0008-5472.CAN-12-3232.

Khoury, A. *et al.* (2020) 'Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains', *Nature Communications*. Nature Research, 11(1), pp. 1–13. doi: 10.1038/s41467-019-13753-7.

Kikuchi, H. *et al.* (2014a) 'GCN5 is essential for IRF-4 gene expression followed by transcriptional activation of Blimp-1 in immature B cells', *Journal of Leukocyte Biology*. Wiley, 95(3), pp. 399–404. doi: 10.1189/jlb.0413232.

Kikuchi, H. *et al.* (2014b) 'GCN5 is involved in regulation of immunoglobulin heavy chain gene expression in immature B cells', *Gene*. Elsevier, 544(1), pp. 19–24. doi: 10.1016/j.gene.2014.04.030.

Kim, K. T. *et al.* (2015) 'Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells', *Genome Biology*. BioMed Central Ltd., 16(1). doi: 10.1186/s13059-015-0692-3.

Kim, S. J. *et al.* (2007) 'Mitochondrial isocitrate dehydrogenase protects human neuroblastoma SH-SY5Y cells against oxidative stress', *Journal of Neuroscience Research*. John Wiley & Sons, Ltd, 85(1), pp. 139–152. doi: 10.1002/jnr.21106.

Kim, T. M. *et al.* (2015) 'Subclonal genomic architectures of primary and metastatic colorectal

cancer based on intratumoral genetic heterogeneity', *Clinical Cancer Research*. American Association for Cancer Research Inc., 21(19), pp. 4461–4472. doi: 10.1158/1078-0432.CCR-14-2413.

Kimura, M. (1983) 'Rare variant alleles in the light of the neutral theory', *Molecular Biology and Evolution*. Mol Biol Evol, 1(1), pp. 84–93. doi: 10.1093/oxfordjournals.molbev.a040305.

Kinsella, R. J. *et al.* (2011) 'Ensembl BioMarts: A hub for data retrieval across taxonomic space', *Database*. Database (Oxford), 2011. doi: 10.1093/database/bar030.

Kivioja, T. *et al.* (2012) 'Counting absolute numbers of molecules using unique molecular identifiers', *Nature Methods*. Nat Methods, 9(1), pp. 72–74. doi: 10.1038/nmeth.1778.

Klco, J. M. *et al.* (2014) 'Functional heterogeneity of genetically defined subclones in acute myeloid leukemia', *Cancer Cell*. Cell Press, 25(3), pp. 379–392. doi: 10.1016/j.ccr.2014.01.031.

Klco, J. M. *et al.* (2015) 'Association between mutation clearance after induction therapy and outcomes in acute myeloid leukemia', *JAMA - Journal of the American Medical Association*. American Medical Association, 314(8), pp. 811–822. doi: 10.1001/jama.2015.9643.

Klein, C. A. *et al.* (2002) 'Combined transcriptome and genome analysis of single micrometastatic cells', *Nature Biotechnology*. Nat Biotechnol, 20(4), pp. 387–392. doi: 10.1038/nbt0402-387.

Klinge, S. and Woolford, J. L. (2019) 'Ribosome assembly coming into focus', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 116–131. doi: 10.1038/s41580-018-0078-y.

Kloosterman, W. P. *et al.* (2011) 'Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer.', *Genome biology*. Genome Biol, 12(10). doi: 10.1186/gb-2011-12-10-r103.

Ko, M. S. H. (1991) 'A stochastic model for gene induction', *Journal of Theoretical Biology*. Academic Press, 153(2), pp. 181–194. doi: 10.1016/S0022-5193(05)80421-7.

Kobayashi, M. *et al.* (2017) 'Phosphatase PRL2 promotes AML1-ETO-induced acute myeloid

leukemia', *Leukemia*. Nature Publishing Group, pp. 1453–1457. doi: 10.1038/leu.2017.67.

Koeffler, H. P. and Leong, G. (2017) 'Preleukemia: One name, many meanings', *Leukemia*. Nature Publishing Group, pp. 534–542. doi: 10.1038/leu.2016.364.

Kogan, S. C. *et al.* (2002) 'Bethesda proposals for classification of nonlymphoid hematopoietic neoplasms in mice', *Blood*, 100(1), pp. 238–245. doi: 10.1182/blood.V100.1.238.

Köhler, A. *et al.* (2008) 'Yeast Ataxin-7 links histone deubiquitination with gene gating and mRNA export', *Nature Cell Biology*. Nature Publishing Group, 10(6), pp. 707–715. doi: 10.1038/ncb1733.

Kondrashov, N. *et al.* (2011a) 'Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning', *Cell*. Cell, 145(3), pp. 383–397. doi: 10.1016/j.cell.2011.03.028.

Kondrashov, N. *et al.* (2011b) 'Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning', *Cell*. NIH Public Access, 145(3), pp. 383–397. doi: 10.1016/j.cell.2011.03.028.

Kostareli, E. *et al.* (2012) 'AML1/ETO and POU4F1 synergy drives B-lymphoid gene expression typical of t(8;21) acute myeloid leukemia', *Leukemia*. doi: 10.1038/leu.2011.316.

Krejci, O. *et al.* (2008) 'P53 signaling in response to increased DNA damage sensitizes AML1-ETO cells to stress-induced death', *Blood*. Blood, 111(4), pp. 2190–2199. doi: 10.1182/blood-2007-06-093682.

Kreso, A. *et al.* (2013a) 'Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer', *Science*. American Association for the Advancement of Science, 339(6119), pp. 543–548. doi: 10.1126/science.1227670.

Kreso, A. *et al.* (2013b) 'Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer', *Science*. American Association for the Advancement of Science, 339(6119), pp. 543–548. doi: 10.1126/science.1227670.

Krivtsov, A. V. and Armstrong, S. A. (2007) 'MLL translocations, histone modifications and leukaemia stem-cell development', *Nature Reviews Cancer*. Nat Rev Cancer, pp. 823–833. doi: 10.1038/nrc2253.

Krönke, J. *et al.* (2013) ‘Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia’, *Blood*. Blood, 122(1), pp. 100–108. doi: 10.1182/blood-2013-01-479188.

Kühn, R. *et al.* (1995a) ‘Inducible gene targeting in mice’, *Science*. American Association for the Advancement of Science, 269(5229), pp. 1427–1429. doi: 10.1126/science.7660125.

Kühn, R. *et al.* (1995b) ‘Inducible gene targeting in mice’, *Science*. American Association for the Advancement of Science, 269(5229), pp. 1427–1429. doi: 10.1126/science.7660125.

Kumar, N., Singh, A. and Kulkarni, R. V. (2015) ‘Transcriptional Bursting in Gene Expression: Analytical Results for General Stochastic Models’, *PLoS Computational Biology*. Public Library of Science, 11(10). doi: 10.1371/journal.pcbi.1004292.

Kurat, C. F. *et al.* (2017) ‘Chromatin Controls DNA Replication Origin Selection, Lagging-Strand Synthesis, and Replication Fork Rates’, *Molecular Cell*. Cell Press, 65(1), pp. 117–130. doi: 10.1016/j.molcel.2016.11.016.

Kurimoto, K. *et al.* (2006) ‘An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis’, *Nucleic Acids Research*. Nucleic Acids Res, 34(5). doi: 10.1093/nar/gkl050.

Lam, F. H., Steger, D. J. and O’Shea, E. K. (2008) ‘Chromatin decouples promoter threshold from dynamic range’, *Nature*. Nature Publishing Group, 453(7192), pp. 246–250. doi: 10.1038/nature06867.

Lamb, R. *et al.* (2015) ‘Antibiotics that target mitochondria effectively eradicate cancer stem cells, across multiple tumor types: Treating cancer like an infectious disease’, *Oncotarget*. Impact Journals LLC, 6(7), pp. 4569–4584. doi: 10.18632/oncotarget.3174.

Landau, D. A. *et al.* (2013) ‘Evolution and impact of subclonal mutations in chronic lymphocytic leukemia’, *Cell*. Elsevier, 152(4), pp. 714–726. doi: 10.1016/j.cell.2013.01.019.

Landau, D. A. *et al.* (2014) ‘Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia’, *Cancer Cell*. Cell Press, 26(6), pp. 813–825. doi: 10.1016/j.ccell.2014.10.012.

Landau, D. A. *et al.* (2015) ‘Mutations driving CLL and their evolution in progression and

relapse', *Nature*. Nature Publishing Group, 526(7574), pp. 525–530. doi: 10.1038/nature15395.

Lang, B. F., Gray, M. W. and Burger, G. (1999) 'Mitochondrial genome evolution and the origin of eukaryotes', *Annual Review of Genetics*. Annu Rev Genet, pp. 351–397. doi: 10.1146/annurev.genet.33.1.351.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*. Nature Publishing Group, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Larsson, A. J. M. *et al.* (2019) 'Genomic encoding of transcriptional burst kinetics', *Nature*. Nature Publishing Group, 565(7738), pp. 251–254. doi: 10.1038/s41586-018-0836-1.

Laurie, C. C. *et al.* (2012) 'Detectable clonal mosaicism from birth to old age and its relationship to cancer', *Nature Genetics*. Nat Genet, 44(6), pp. 642–650. doi: 10.1038/ng.2271.

Lavau, C. *et al.* (1997) *Immortalization and leukemic transformation of a myelomonocytic precursor by retrovirally transduced HRX-ENL blastic leukemias (ALL) and myelomonocytic and mono*, *The EMBO Journal*.

Lawrence, M. S. *et al.* (2013a) 'Mutational heterogeneity in cancer and the search for new cancer-associated genes', *Nature*. Nature Publishing Group, 499(7457), pp. 214–218. doi: 10.1038/nature12213.

Lawrence, M. S. *et al.* (2013b) 'Mutational heterogeneity in cancer and the search for new cancer-associated genes', *Nature*. NIH Public Access, 499(7457), pp. 214–218. doi: 10.1038/nature12213.

Lawrence, M. S. *et al.* (2014) 'Discovery and saturation analysis of cancer genes across 21 tumour types', *Nature*. Nature, 505(7484), pp. 495–501. doi: 10.1038/nature12912.

Lee, K. K. *et al.* (2011) 'Combinatorial depletion analysis to assemble the network architecture of the SAGA and ADA chromatin remodeling complexes', *Molecular Systems Biology*. John Wiley & Sons, Ltd, 7(1), p. 503. doi: 10.1038/msb.2011.40.

Lee, S. *et al.* (2006) 'Gene expression profiles in acute myeloid leukemia with common translocations using SAGE', *Proceedings of the National Academy of Sciences of the United*

States of America. National Academy of Sciences, 103(4), pp. 1030–1035. doi: 10.1073/pnas.0509878103.

Lee, S. M. *et al.* (2002) 'Cytosolic NADP+-dependent isocitrate dehydrogenase status modulates oxidative damage to cells', *Free Radical Biology and Medicine*. Free Radic Biol Med, 32(11), pp. 1185–1196. doi: 10.1016/S0891-5849(02)00815-8.

Leek, J. T. *et al.* (2010) 'Tackling the widespread and critical impact of batch effects in high-throughput data', *Nature Reviews Genetics*. Nat Rev Genet, pp. 733–739. doi: 10.1038/nrg2825.

Lemasters, J. J. *et al.* (1998) 'The mitochondrial permeability transition in cell death: A common mechanism in necrosis, apoptosis and autophagy', *Biochimica et Biophysica Acta - Bioenergetics*. Elsevier, 1366(1–2), pp. 177–196. doi: 10.1016/S0005-2728(98)00112-1.

Levsky, J. M. *et al.* (2002) 'Single-cell gene expression profiling', *Science*. Science, 297(5582), pp. 836–840. doi: 10.1126/science.1072241.

Li, S. *et al.* (2014) 'Dynamic evolution of clonal epialleles revealed by methclone', *Genome Biology*. BioMed Central Ltd., 15(9), p. 472. doi: 10.1186/s13059-014-0472-5.

Li, S. *et al.* (2016) 'Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia', *Nature Medicine*. Nature Publishing Group, 22(7), pp. 792–799. doi: 10.1038/nm.4125.

Liberzon, A. *et al.* (2015) 'The Molecular Signatures Database Hallmark Gene Set Collection', *Cell Systems*. Cell Press, 1(6), pp. 417–425. doi: 10.1016/j.cels.2015.12.004.

Lim, H. N. and Van Oudenaarden, A. (2007) 'A multistep epigenetic switch enables the stable inheritance of DNA methylation states', *Nature Genetics*. Nat Genet, 39(2), pp. 269–275. doi: 10.1038/ng1956.

Lin, W. *et al.* (2008) 'Proper expression of the Gcn5 histone acetyltransferase is required for neural tube closure in mouse embryos', *Developmental Dynamics*. John Wiley & Sons, Ltd, 237(4), pp. 928–940. doi: 10.1002/dvdy.21479.

Linder, D. and Gartler, S. M. (1965) 'Glucose-6-phosphate dehydrogenase mosaicism:

Utilization as a cell marker in the study of leiomyomas', *Science*. Science, 150(3692), pp. 67–69. doi: 10.1126/science.150.3692.67.

Ling, S. *et al.* (2015) 'Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 112(47), pp. E6496–E6505. doi: 10.1073/pnas.1519556112.

Lippman, Z. *et al.* (2004) 'Role of transposable elements in heterochromatin and epigenetic control', *Nature*. Nature, 430(6998), pp. 471–476. doi: 10.1038/nature02651.

Liu, T. *et al.* (2006) 'Histone deacetylase inhibitors: Multifunctional anticancer agents', *Cancer Treatment Reviews*, 32(3), pp. 157–165. doi: 10.1016/j.ctrv.2005.12.006.

Liu, Y. *et al.* (2007) 'Structural Basis for Recognition of SMRT/N-CoR by the MYND Domain and Its Contribution to AML1/ETO's Activity', *Cancer Cell*, 11(6), pp. 483–497. doi: 10.1016/j.ccr.2007.04.010.

Liu, Y. *et al.* (2017) 'The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia', *Nature Genetics*. Nature Publishing Group, 49(8), pp. 1211–1218. doi: 10.1038/ng.3909.

Liu, Z. *et al.* (2011) 'Control of embryonic stem cell lineage commitment by core promoter factor, TAF3', *Cell*. Elsevier, 146(5), pp. 720–731. doi: 10.1016/j.cell.2011.08.005.

Ljungström, V. *et al.* (2016a) 'Whole-exome sequencing in relapsing chronic lymphocytic leukemia: Clinical impact of recurrent RPS15 mutations', *Blood*. American Society of Hematology, 127(8), pp. 1007–1016. doi: 10.1182/blood-2015-10-674572.

Ljungström, V. *et al.* (2016b) 'Whole-exome sequencing in relapsing chronic lymphocytic leukemia: Clinical impact of recurrent RPS15 mutations', *Blood*. American Society of Hematology, 127(8), pp. 1007–1016. doi: 10.1182/blood-2015-10-674572.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*. BioMed Central Ltd., 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lowenberg, B., Downing, J. R. and Burnett, A. (1999) 'Acute Myeloid Leukemia', *New England Journal of Medicine*. Massachusetts Medical Society , 341(14), pp. 1051–1062. doi: 10.1056/NEJM199909303411407.

Lüer, K. and Technau, G. M. (2009) 'Single cell cultures of drosophila neuroectodermal and mesectodermal central nervous system progenitors reveal different degrees of developmental autonomy', *Neural Development*. Neural Dev, 4(1). doi: 10.1186/1749-8104-4-30.

Lynch, M. (2010) 'Rate, molecular spectrum, and consequences of human mutation', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 107(3), pp. 961–968. doi: 10.1073/pnas.0912629107.

Van Der Maaten, L. and Hinton, G. (2008) *Visualizing Data using t-SNE*, *Journal of Machine Learning Research*.

MacArthur, B. D., Maayan, A. and Lemischka, I. R. (2009) 'Systems biology of stem cell fate and cellular reprogramming', *Nature Reviews Molecular Cell Biology*. Nat Rev Mol Cell Biol, pp. 672–681. doi: 10.1038/nrm2766.

Macaulay, I. C. *et al.* (2015) 'G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes', *Nature Methods*. Nature Publishing Group, 12(6), pp. 519–522. doi: 10.1038/nmeth.3370.

Magee, J. A. and Signer, R. A. J. (2021) 'Developmental Stage-Specific Changes in Protein Synthesis Differentially Sensitize Hematopoietic Stem Cells and Erythroid Progenitors to Impaired Ribosome Biogenesis', *Stem Cell Reports*. Cell Press, 16(1), pp. 20–28. doi: 10.1016/j.stemcr.2020.11.017.

Majeti, R. *et al.* (2009) 'Dysregulated gene expression networks in human acute myelogenous leukemia stem cells', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 106(9), pp. 3396–3401. doi: 10.1073/pnas.0900089106.

Mardis, E. R. *et al.* (2009) 'Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 361(11), pp. 1058–1066. doi: 10.1056/nejmoa0903840.

Markman, B., Dienstmann, R. and Tabernero, J. (2010) 'Targeting the PI3K/Akt/mTOR pathway--beyond rapalogs.', *Oncotarget*. Oncotarget, pp. 530–543. doi: 10.18632/oncotarget.188.

Marsin, A. S. *et al.* (2000) 'Phosphorylation and activation of heart PFK-2 by AMPK has a role in the stimulation of glycolysis during ischaemia', *Current Biology*. Current Biology Ltd, 10(20), pp. 1247–1255. doi: 10.1016/S0960-9822(00)00742-9.

Marsin, A. S. *et al.* (2002) 'The stimulation of glycolysis by hypoxia in activated monocytes is mediated by AMP-activated protein kinase and inducible 6-phosphofructo-2-kinase', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 277(34), pp. 30778–30783. doi: 10.1074/jbc.M205213200.

Martincorena, I. and Campbell, P. J. (2015) 'Somatic mutation in cancer and normal cells', *Science*. American Association for the Advancement of Science, pp. 1483–1489. doi: 10.1126/science.aab4082.

Martinez-Soria, N. *et al.* (2018) 'The Oncogenic Transcription Factor RUNX1/ETO Corrupts Cell Cycle Regulation to Drive Leukemic Transformation', *Cancer Cell*. Cell Press, 34(4), pp. 626-642.e8. doi: 10.1016/j.ccell.2018.08.015.

Martinez Arias, A. and Brickman, J. M. (2011) 'Gene expression heterogeneities in embryonic stem cell populations: Origin and function', *Current Opinion in Cell Biology*. Curr Opin Cell Biol, pp. 650–656. doi: 10.1016/j.ceb.2011.09.007.

Martinez, E. *et al.* (2001) 'Human STAGA Complex Is a Chromatin-Acetylating Transcription Coactivator That Interacts with Pre-mRNA Splicing and DNA Damage-Binding Factors In Vivo', *Molecular and Cellular Biology*. American Society for Microbiology, 21(20), pp. 6782–6795. doi: 10.1128/mcb.21.20.6782-6795.2001.

Marusyk, A. and Polyak, K. (2010) 'Tumor heterogeneity: Causes and consequences', *Biochimica et Biophysica Acta - Reviews on Cancer*. Biochim Biophys Acta, pp. 105–117. doi: 10.1016/j.bbcan.2009.11.002.

Mattes, K., Vellenga, E. and Schepers, H. (2019) 'Differential redox-regulation and mitochondrial dynamics in normal and leukemic hematopoietic stem cells: A potential window

for leukemia therapy', *Critical Reviews in Oncology/Hematology*. Elsevier Ireland Ltd, p. 102814. doi: 10.1016/j.critrevonc.2019.102814.

Mcadams, H. H. and Arkin, A. (1997) 'Stochastic mechanisms in gene expression', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 94(3), pp. 814–819. doi: 10.1073/pnas.94.3.814.

McAdams, H. H. and Arkin, A. (1999) 'It's a noisy business! Genetic regulation at the nanomolar scale', *Trends in Genetics*. Trends Genet, pp. 65–69. doi: 10.1016/S0168-9525(98)01659-X.

McGowan, K. A. *et al.* (2011) 'Reduced ribosomal protein gene dosage and p53 activation in low-risk myelodysplastic syndrome', *Blood*. Blood, 118(13), pp. 3622–3633. doi: 10.1182/blood-2010-11-318584.

McGranahan, N. *et al.* (2012) 'Cancer chromosomal instability: Therapeutic and diagnostic challenges', *EMBO Reports*. EMBO Rep, pp. 528–538. doi: 10.1038/embor.2012.61.

Mclean, C. Y. *et al.* (2010) 'A n A ly s i s GREAT improves functional interpretation of cis-regulatory regions', *nature biotechnology VOLUME*, 28. doi: 10.1038/nbt.1630.

McMahon, S. B. *et al.* (1998) 'The novel ATM-related protein TRRAP is an essential cofactor for the c- Myc and E2F oncoproteins', *Cell*. Cell Press, 94(3), pp. 363–374. doi: 10.1016/S0092-8674(00)81479-8.

McPherson, A. *et al.* (2016) 'Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer', *Nature Genetics*. Nature Publishing Group, 48(7), pp. 758–767. doi: 10.1038/ng.3573.

Merkenschlager, M. and Nora, E. P. (2016) 'CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation', *Annual Review of Genomics and Human Genetics*. Annual Reviews Inc., 17(1), pp. 17–43. doi: 10.1146/annurev-genom-083115-022339.

Meyer, C. *et al.* (2009) 'New insights to the MLL recombinome of acute leukemias', *Leukemia*. Nature Publishing Group, 23(8), pp. 1490–1499. doi: 10.1038/leu.2009.33.

Meyer, C. *et al.* (2013) 'The MLL recombinome of acute leukemias in 2013', *Leukemia*.

Leukemia, 27(11), pp. 2165–2176. doi: 10.1038/leu.2013.135.

Mezger, A. *et al.* (2018) 'High-throughput chromatin accessibility profiling at single-cell resolution', *Nature Communications*. Nature Publishing Group, 9(1). doi: 10.1038/s41467-018-05887-x.

Mi, H. *et al.* (2018) 'PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools', *Nucleic Acids Research*, 47, pp. 419–426. doi: 10.1093/nar/gky1038.

Mi, R. K. *et al.* (2009) 'Mutational analysis of IDH1 codon 132 in glioblastomas and other common cancers', *International Journal of Cancer*. Int J Cancer, 125(2), pp. 353–355. doi: 10.1002/ijc.24379.

Mi, W. *et al.* (2017) 'YEATS2 links histone acetylation to tumorigenesis of non-small cell lung cancer', *Nature Communications*. Nature Publishing Group, 8(1). doi: 10.1038/s41467-017-01173-4.

Mi, W. *et al.* (2018) 'The ZZ-type zinc finger of ZZZ3 modulates the ATAC complex-mediated histone acetylation and gene activation', *Nature Communications*. Nature Publishing Group, 9(1). doi: 10.1038/s41467-018-06247-5.

Miller-Jensen, K. *et al.* (2011) 'Varying virulence: Epigenetic control of expression noise and disease processes', *Trends in Biotechnology*. Trends Biotechnol, pp. 517–525. doi: 10.1016/j.tibtech.2011.05.004.

Mills, E. W. and Green, R. (2017) 'Ribosomopathies: There's strength in numbers', *Science*. American Association for the Advancement of Science. doi: 10.1126/science.aan2755.

Milne, T. A. *et al.* (2002) 'MLL targets SET domain methyltransferase activity to Hox gene promoters', *Molecular Cell*. Cell Press, 10(5), pp. 1107–1117. doi: 10.1016/S1097-2765(02)00741-4.

Miyoshi, H. *et al.* (1991) '(8;21) breakpoints on chromosome 21 in acute myeloid leukemia are clustered within a limited region of a single gene, AML1', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 88(23),

pp. 10431–10434. doi: 10.1073/pnas.88.23.10431.

Miyoshi, H. *et al.* (1993) 'The t(8;21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript.', *The EMBO Journal*. John Wiley & Sons, Ltd, 12(7), pp. 2715–2721. doi: 10.1002/j.1460-2075.1993.tb05933.x.

Mohammed, H. *et al.* (2017) 'Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation', *CellReports*, 20, pp. 1215–1228. doi: 10.1016/j.celrep.2017.07.009.

Moignard, V. and Göttgens, B. (2014) 'Transcriptional mechanisms of cell fate decisions revealed by single cell expression profiling', *BioEssays*. Wiley-Blackwell, 36(4), pp. 419–426. doi: 10.1002/bies.201300102.

Molavi, G., Samadi, N. and Hosseingholi, E. Z. (2019) 'The roles of moonlight ribosomal proteins in the development of human cancers', *Journal of Cellular Physiology*. Wiley-Liss Inc., pp. 8327–8341. doi: 10.1002/jcp.27722.

Mori, H. *et al.* (2002) 'Chromosome translocations and covert leukemic clones are generated during normal fetal development', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 99(12), pp. 8242–8247. doi: 10.1073/pnas.112218799.

Moris, N. *et al.* (2018a) 'Histone Acetyltransferase KAT2A Stabilizes Pluripotency with Control of Transcriptional Heterogeneity', *Stem Cells*. Wiley-Blackwell, 36(12), pp. 1828–1838. doi: 10.1002/stem.2919.

Moris, N. *et al.* (2018b) 'Histone Acetyltransferase KAT2A Stabilizes Pluripotency with Control of Transcriptional Heterogeneity', *STEM CELLS*. Wiley-Blackwell, 36(12), pp. 1828–1838. doi: 10.1002/stem.2919.

Moris, N., Pina, C. and Arias, A. M. (2016a) 'Transition states and cell fate decisions in epigenetic landscapes', *Nature Reviews Genetics*. Nature Publishing Group, pp. 693–703. doi: 10.1038/nrg.2016.98.

Moris, N., Pina, C. and Arias, A. M. (2016b) 'Transition states and cell fate decisions in

epigenetic landscapes', *Nature Reviews Genetics*. Nature Publishing Group, pp. 693–703. doi: 10.1038/nrg.2016.98.

Morris, J., Singh, J. M. and Eberwine, J. H. (2011) 'Transcriptome analysis of single cells', *Journal of Visualized Experiments*. Journal of Visualized Experiments, (50). doi: 10.3791/2634.

Moulton, T. *et al.* (1996) 'Genomic imprinting and Wilms' tumor', *Medical and Pediatric Oncology*. Med Pediatr Oncol, 27(5), pp. 476–483. doi: 10.1002/(SICI)1096-911X(199611)27:5<476::AID-MPO15>3.0.CO;2-8.

Müller, S. *et al.* (2016) ' Single-cell sequencing maps gene expression to mutational phylogenies in PDGF - and EGF -driven gliomas ', *Molecular Systems Biology*. EMBO, 12(11), p. 889. doi: 10.15252/msb.20166969.

Nafria, M. *et al.* (2020) 'Expression of RUNX1-ETO Rapidly Alters the Chromatin Landscape and Growth of Early Human Myeloid Precursor Cells', *Cell Reports*. Elsevier B.V., 31(8), p. 107691. doi: 10.1016/j.celrep.2020.107691.

Narla, A. and Ebert, B. L. (2010) 'Ribosomopathies: Human disorders of ribosome dysfunction', *Blood*. Blood, pp. 3196–3205. doi: 10.1182/blood-2009-10-178129.

Neuwald, A. F. and Landsman, D. (1997) 'GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein', *Trends in Biochemical Sciences*. Elsevier Ltd, pp. 154–155. doi: 10.1016/S0968-0004(97)01034-7.

Nicolas, D. *et al.* (2018) 'Modulation of transcriptional burst frequency by histone acetylation', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(27), pp. 7153–7158. doi: 10.1073/pnas.1722330115.

Nik-Zainal, S., Alexandrov, L. B., *et al.* (2012) 'Mutational processes molding the genomes of 21 breast cancers', *Cell*. Elsevier, 149(5), pp. 979–993. doi: 10.1016/j.cell.2012.04.024.

Nik-Zainal, S., Van Loo, P., *et al.* (2012) 'The life history of 21 breast cancers', *Cell*. Cell Press, 149(5), pp. 994–1007. doi: 10.1016/j.cell.2012.04.023.

Nikolaev, S. I. *et al.* (2012) 'A single-nucleotide substitution mutator phenotype revealed by

exome sequencing of human colon adenomas', *Cancer Research*. American Association for Cancer Research, 72(23), pp. 6279–6289. doi: 10.1158/0008-5472.CAN-12-3869.

Nilsson, L. *et al.* (2007) 'The molecular signature of MDS stem cells supports a stem-cell origin of 5q-myelodysplastic syndromes', *Blood*. Blood, 110(8), pp. 3005–3014. doi: 10.1182/blood-2007-03-079368.

Nora, E. P. *et al.* (2020) 'Molecular basis of CTCF binding polarity in genome folding', *Nature Communications*. Nature Research, 11(1). doi: 10.1038/s41467-020-19283-x.

Notta, F. *et al.* (2016) 'Distinct routes of lineage development reshape the human blood hierarchy across ontogeny', *Science*. American Association for the Advancement of Science, 351(6269). doi: 10.1126/science.aab2116.

Novick, A. and Weiner, M. (1957) 'ENZYME INDUCTION AS AN ALL-OR-NONE PHENOMENON', *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences, 43(7), pp. 553–566. doi: 10.1073/pnas.43.7.553.

Nowell, P. C. (1976) *The Clonal Evolution of Tumor Cell Populations*, *New Series*.

O'Brien, T. W. (2003) 'Properties of Human Mitochondrial Ribosomes', *IUBMB Life*. IUBMB Life, pp. 505–513. doi: 10.1080/15216540310001626610.

Ochiai, H. *et al.* (2020) 'Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells', *Science Advances*. American Association for the Advancement of Science, 6(25), pp. 6699–6716. doi: 10.1126/sciadv.aaz6699.

Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001) 'CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease', *Trends in Genetics*. Elsevier, pp. 520–527. doi: 10.1016/S0168-9525(01)02366-6.

Osato, M. *et al.* (1999) 'Biallelic and heterozygous point mutations in the runt domain of the AML1/PEBP2αB gene associated with myeloblastic leukemias', *Blood*. W.B. Saunders, 93(6), pp. 1817–1824. doi: 10.1182/blood.v93.6.1817.406k36_1817_1824.

Ozbudak, E. M. *et al.* (2002) 'Regulation of noise in the expression of a single gene', *Nature Genetics*. Nature Publishing Group, 31(1), pp. 69–73. doi: 10.1038/ng869.

P.Greenberg (1983) ‘The smouldering myeloid leukaemic states’, *Blood*, 61(103).

Paguirigan, A. L. *et al.* (2015) ‘Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia’, *Science Translational Medicine*. American Association for the Advancement of Science, 7(281), p. 281re2. doi: 10.1126/scitranslmed.aaa0763.

Pan, H. *et al.* (2015) ‘Epigenomic evolution in diffuse large B-cell lymphomas’, *Nature Communications*. Nature Publishing Group, 6(1), pp. 1–12. doi: 10.1038/ncomms7921.

Papaemmanuil, E. *et al.* (2016) ‘Genomic Classification and Prognosis in Acute Myeloid Leukemia’, *The New England journal of medicine*. Europe PMC Funders, 374(23), p. 2209. doi: 10.1056/NEJMOA1516192.

Paulsson, J., Berg, O. G. and Ehrenberg, M. (2000) ‘Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 97(13), pp. 7148–7153. doi: 10.1073/pnas.110057697.

Pavlides, S. *et al.* (2009) ‘The reverse Warburg effect: Aerobic glycolysis in cancer associated fibroblasts and the tumor stroma’, *Cell Cycle*. Taylor and Francis Inc., 8(23), pp. 3984–4001. doi: 10.4161/cc.8.23.10238.

Pearce, L. R. *et al.* (2010) ‘Characterization of PF-4708671, a novel and highly specific inhibitor of p70 ribosomal S6 kinase (S6K1)’, *Biochemical Journal*. Biochem J, 431(2), pp. 245–255. doi: 10.1042/BJ20101024.

Peccoud, J. and Ycart, B. (1995) ‘Markovian modeling of gene-product synthesis’, *Theoretical Population Biology*. Academic Press, 48(2), pp. 222–234. doi: 10.1006/tpbi.1995.1027.

Pellagatti, A. *et al.* (2008) ‘Haploinsufficiency of RPS14 in 5q- syndrome is associated with deregulation of ribosomal- and translation-related genes’, *British Journal of Haematology*. Br J Haematol, 142(1), pp. 57–64. doi: 10.1111/j.1365-2141.2008.07178.x.

Peterson, L. F. and Zhang, D. E. (2004) ‘The 8;21 translocation in leukemogenesis’, *Oncogene*. Oncogene, pp. 4255–4262. doi: 10.1038/sj.onc.1207727.

Pfaffl, M. W. (2001) ‘A new mathematical model for relative quantification in real-time RT-

PCR.', *Nucleic acids research*. Oxford University Press, 29(9), p. e45. doi: 10.1093/nar/29.9.e45.

Pfeifer, G. P. *et al.* (2002) 'Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers', *Oncogene*. Oncogene, 21–48(6), pp. 7435–7451. doi: 10.1038/sj.onc.1205803.

Pina, C. *et al.* (2012) 'Inferring rules of lineage commitment in haematopoiesis', *Nature Cell Biology*. Nat Cell Biol, 14(3), pp. 287–294. doi: 10.1038/ncb2442.

Pina, C. and Enver, T. (2007) 'Differential contributions of haematopoietic stem cells to foetal and adult haematopoiesis: Insights from functional analysis of transcriptional regulators', *Oncogene*. Oncogene, pp. 6750–6765. doi: 10.1038/sj.onc.1210759.

Ptasinska, A. *et al.* (2012) 'Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding', *Leukemia*. Nature Publishing Group, 26(8), pp. 1829–1841. doi: 10.1038/leu.2012.49.

Ptasinska, A. *et al.* (2014) 'Identification of a dynamic core transcriptional network in t(8;21) AML that regulates differentiation block and self-renewal', *Cell Reports*. Elsevier, 8(6), pp. 1974–1988. doi: 10.1016/j.celrep.2014.08.024.

Ptasinska, A. *et al.* (2019) 'RUNX1-ETO Depletion in t(8;21) AML Leads to C/EBPα- and AP-1-Mediated Alterations in Enhancer-Promoter Interaction', *Cell Reports*. Elsevier B.V., 28(12), pp. 3022-3031.e7. doi: 10.1016/j.celrep.2019.08.040.

Van de Putte, P. and Goosen, N. (1992) 'DNA inversions in phages and bacteria', *Trends in Genetics*. Trends Genet, pp. 457–462. doi: 10.1016/0168-9525(92)90331-W.

Qiu, P. (2020) 'Embracing the dropouts in single-cell RNA-seq analysis', *Nature Communications*. Nature Research, 11(1). doi: 10.1038/s41467-020-14976-9.

Quentin, S. *et al.* (2011) 'Myelodysplasia and leukemia of fanconi anemia are associated with a specific pattern of genomic abnormalities that includes cryptic RUNX1/AML1 lesions', *Blood*. American Society of Hematology, 117(15), pp. e161–e170. doi: 10.1182/blood-2010-09-308726.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *BIOINFORMATICS APPLICATIONS NOTE*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Raj, A. *et al.* (2006) 'Stochastic mRNA synthesis in mammalian cells', *PLoS Biology*. PLoS Biol, 4(10), pp. 1707–1719. doi: 10.1371/journal.pbio.0040309.

Raj, A. *et al.* (2008) 'Imaging individual mRNA molecules using multiple singly labeled probes', *Nature Methods*. Nat Methods, 5(10), pp. 877–879. doi: 10.1038/nmeth.1253.

Raj, A. and van Oudenaarden, A. (2008) 'Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences', *Cell*. NIH Public Access, pp. 216–226. doi: 10.1016/j.cell.2008.09.050.

Raj, A. and Van Oudenaarden, A. (2009) 'Single-molecule approaches to stochastic gene expression', *Annual Review of Biophysics*. NIH Public Access, pp. 255–270. doi: 10.1146/annurev.biophys.37.032807.125928.

Raser, J. M. and O'Shea, E. K. (2004a) 'Control of stochasticity in eukaryotic gene expression', *Science*. Science, 304(5678), pp. 1811–1814. doi: 10.1126/science.1098641.

Raser, J. M. and O'Shea, E. K. (2004b) 'Control of stochasticity in eukaryotic gene expression', *Science*. American Association for the Advancement of Science, 304(5678), pp. 1811–1814. doi: 10.1126/science.1098641.

Rashkovan, M. and Ferrando, A. (2019) 'Metabolic dependencies and vulnerabilities in leukemia', *Genes & development*. NLM (Medline), pp. 1460–1474. doi: 10.1101/gad.326470.119.

Rasmussen, K. D. *et al.* (2015) 'Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis'. doi: 10.1101/gad.260174.115.

Raval, A. *et al.* (2012) 'Reduced rRNA expression and increased rDNA promoter methylation in CD34+ cells of patients with myelodysplastic syndromes', *Blood*. The American Society of Hematology, 120(24), pp. 4812–4818. doi: 10.1182/blood-2012-04-423111.

Rebel, V. I. *et al.* (2002) 'Distinct roles for CREB-binding protein and p300 in hematopoietic stem cell self-renewal', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 99(23), pp. 14789–14794. doi: 10.1073/pnas.232568499.

Redondo Monte, E. *et al.* (2020) 'ZBTB7A prevents RUNX1-RUNX1T1-dependent clonal expansion of human hematopoietic stem and progenitor cells', *Oncogene*. Springer Nature, 39(15), pp. 3195–3205. doi: 10.1038/s41388-020-1209-4.

Regev, A. *et al.* (2017) 'The human cell atlas', *eLife*. eLife Sciences Publications Ltd, 6. doi: 10.7554/eLife.27041.

Ren, G. *et al.* (2017a) 'CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression', *Molecular Cell*. Cell Press, 67(6), pp. 1049-1058.e6. doi: 10.1016/j.molcel.2017.08.026.

Ren, G. *et al.* (2017b) 'CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression', *Molecular Cell*. Cell Press, 67(6), pp. 1049-1058.e6. doi: 10.1016/j.molcel.2017.08.026.

Ren, G. and Zhao, K. (2019) 'CTCF and cellular heterogeneity', *Cell and Bioscience*. BioMed Central Ltd., pp. 1–7. doi: 10.1186/s13578-019-0347-2.

Rhoades, K. L. *et al.* (2000) *Analysis of the role of AML1-ETO in leukemogenesis, using an inducible transgenic mouse model*. Available at: http://ashpublications.org/blood/article-pdf/96/6/2108/1667636/h8180002108.pdf (Accessed: 6 December 2020).

Ribeiro, A. S. *et al.* (2010) 'Dynamical effects of transcriptional pause-prone sites', *Computational Biology and Chemistry*. Elsevier, 34(3), pp. 143–148. doi: 10.1016/j.compbiolchem.2010.04.003.

Richards, E. J. (2006) 'Inherited epigenetic variation - Revisiting soft inheritance', *Nature Reviews Genetics*. Nat Rev Genet, pp. 395–401. doi: 10.1038/nrg1834.

Van Riggelen, J., Yetil, A. and Felsher, D. W. (2010) 'MYC as a regulator of ribosome biogenesis and protein synthesis', *Nature Reviews Cancer*. Nat Rev Cancer, pp. 301–309. doi:

10.1038/nrc2819.

Rivlin, N. *et al.* (2011) ‘Mutations in the p53 tumor suppressor gene: Important milestones at the various steps of tumorigenesis’, *Genes and Cancer*. Impact Journals, LLC, pp. 466–474. doi: 10.1177/1947601911408889.

RJ, J. and C, D. (2010) ‘Stochastic mechanisms of cell fate specification that yield random or robust outcomes’, *Annual review of cell and developmental biology*. Annu Rev Cell Dev Biol, 26. doi: 10.1146/ANNUREV-CELLBIO-100109-104113.

Robb, M. L. and Shahrezaei, V. (2014) ‘Stochastic cellular fate decision making by multiple infecting lambda phage’, *PLoS ONE*. Public Library of Science, 9(8). doi: 10.1371/journal.pone.0103636.

Rosen, N. and She, Q. B. (2006) ‘AKT and cancer-Is it all mTOR?’, *Cancer Cell*. Cancer Cell, pp. 254–256. doi: 10.1016/j.ccr.2006.10.001.

Rosenfeld, N. *et al.* (2005) *Gene regulation at the single-cell level*, *Science*. doi: 10.1126/science.1106914.

Roth, G. A. *et al.* (2018) ‘Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017’, *The Lancet*. Lancet Publishing Group, 392(10159), pp. 1736–1788. doi: 10.1016/S0140-6736(18)32203-7.

Rowley, J. D. (1973) ‘Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia’, *Ann Genet.*, 16(2), pp. 109–12.

Ruggero, D. *et al.* (2004) ‘The translation factor eIF-4E promotes tumor formation and cooperates with c-Myc in lymphomagenesis’, *Nature Medicine*. Nat Med, 10(5), pp. 484–486. doi: 10.1038/nm1042.

Saada, A. *et al.* (2007) ‘Antenatal mitochondrial disease caused by mitochondrial ribosomal protein (MRPS22) mutation’, *Journal of Medical Genetics*. J Med Genet, 44(12), pp. 784–786. doi: 10.1136/jmg.2007.053116.

Saito, Y. *et al.* (2015) ‘AMPK Protects Leukemia-Initiating Cells in Myeloid Leukemias from

Metabolic Stress in the Bone Marrow', *Cell Stem Cell*. Cell Press, 17(5), pp. 585–596. doi: 10.1016/j.stem.2015.08.019.

Salmond, R. J. *et al.* (2015) 'Mechanistic Target of Rapamycin Complex 1/S6 Kinase 1 Signals Influence T Cell Activation Independently of Ribosomal Protein S6 Phosphorylation', *The Journal of Immunology Author Choice*. The American Association of Immunologists, Inc., 195(10), p. 4615. doi: 10.4049/JIMMUNOL.1501473.

Salunkhe, S. *et al.* (2018) 'Inhibition of novel GCN5-ATM axis restricts the onset of acquired drug resistance in leukemia', *International Journal of Cancer*. Wiley-Liss Inc., 142(10), pp. 2175–2185. doi: 10.1002/ijc.31242.

Sanchez, A., Choubey, S. and Kondev, J. (2013) 'Regulation of noise in gene expression', *Annual Review of Biophysics*. Annu Rev Biophys, 42(1), pp. 469–491. doi: 10.1146/annurev-biophys-083012-130401.

Sanyal, A. *et al.* (2012) 'The long-range interaction landscape of gene promoters', *Nature*. Nature Publishing Group, 489(7414), pp. 109–113. doi: 10.1038/nature11279.

Sasagawa, Y. *et al.* (2013) 'Quartz-Seq: A highly reproducible and sensitive single-cell RNA sequencing method, reveals nongenetic gene-expression heterogeneity', *Genome Biology*. BioMed Central, 14(4). doi: 10.1186/gb-2013-14-4-r31.

Sasaki, M. *et al.* (2012a) 'IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics', *Nature*. Nature, 488(7413), pp. 656–659. doi: 10.1038/nature11323.

Sasaki, M. *et al.* (2012b) 'IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics', *Nature*. Nature Publishing Group, 488(7413), pp. 656–659. doi: 10.1038/nature11323.

Sbarrato, T. *et al.* (2016) 'A ribosome-related signature in peripheral blood CLL B cells is linked to reduced survival following treatment', *Cell Death and Disease*. Nature Publishing Group, 7(6). doi: 10.1038/cddis.2016.148.

Scaduto, R. C. and Grotyohann, L. W. (1999) 'Measurement of mitochondrial membrane

potential using fluorescent rhodamine derivatives', *Biophysical Journal*. Biophysical Society, 76(1 I), pp. 469–477. doi: 10.1016/S0006-3495(99)77214-0.

Schep, A. N. *et al.* (2017) 'ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data', *Nature Methods*. Nature Publishing Group, 14(10), pp. 975–978. doi: 10.1038/nmeth.4401.

Schnittger, S. *et al.* (2011) 'RUNX1 mutations are frequent in de novo AML with noncomplex karyotype and confer an unfavorable prognosis', *Blood*. Blood, 117(8), pp. 2348–2357. doi: 10.1182/blood-2009-11-255976.

Schröter, C. *et al.* (2015) 'FGF/MAPK signaling sets the switching threshold of a bistable circuit controlling cell fate decisions in embryonic stem cells', *Development (Cambridge)*. Company of Biologists Ltd, 142(24), pp. 4205–4216. doi: 10.1242/dev.127530.

SH, U., D, D. and G, T. (2006) 'Nutrient overload, insulin resistance, and ribosomal protein S6 kinase 1, S6K1', *Cell metabolism*. Cell Metab, 3(6). doi: 10.1016/J.CMET.2006.05.003.

Shaffer, S. M. *et al.* (2017) 'Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance', *Nature*. Nature Publishing Group, 546(7658), pp. 431–435. doi: 10.1038/nature22794.

Shah, S. N., Hile, S. E. and Eckert, K. A. (2010) 'Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes', *Cancer Research*. Cancer Res, pp. 431–435. doi: 10.1158/0008-5472.CAN-09-3049.

Shah, S. P. *et al.* (2009) 'Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution', *Nature*. Nature Publishing Group, 461(7265), pp. 809–813. doi: 10.1038/nature08489.

Shah, S. P. *et al.* (2012) 'The clonal and mutational evolution spectrum of primary triple-negative breast cancers', *Nature*. Nature, 486(7403), pp. 395–399. doi: 10.1038/nature10933.

Shih, A. H. *et al.* (2015) 'Mutational cooperativity linked to combinatorial epigenetic gain of function in acute myeloid leukemia', *Cancer Cell*. Cell Press, 27(4), pp. 502–515. doi: 10.1016/j.ccell.2015.03.009.

Shimamura, A. and Alter, B. P. (2010) 'Pathophysiology and management of inherited bone marrow failure syndromes', *Blood Reviews*. Blood Rev, pp. 101–122. doi: 10.1016/j.blre.2010.03.002.

Shlush, L. I. *et al.* (2014a) 'Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia', *Nature*. Nature Publishing Group, 506(7488), pp. 328–333. doi: 10.1038/nature13038.

Shlush, L. I. *et al.* (2014b) 'Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia', *Nature*. Nature Publishing Group, 506(7488), pp. 328–333. doi: 10.1038/nature13038.

Shlush, L. I. *et al.* (2014c) 'Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia', *Nature*. Nature Publishing Group, 506(7488), pp. 328–333. doi: 10.1038/nature13038.

Sieranoja, S. (2018) 'Fast random pair divisive construction of kNN graph using generic distance measures', *ACM*.

Signer, R. A. J. *et al.* (2014) 'Haematopoietic stem cells require a highly regulated protein synthesis rate', *Nature*. Nature Publishing Group, 508(7498), pp. 49–54. doi: 10.1038/nature13035.

Signer, R. A. J. *et al.* (2016) 'The rate of protein synthesis in hematopoietic stem cells is limited partly by 4E-BPs', *Genes and Development*. Cold Spring Harbor Laboratory Press, 30(15), pp. 1698–1703. doi: 10.1101/gad.282756.116.

Simsek, T. *et al.* (2010) 'The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche', *Cell Stem Cell*. NIH Public Access, 7(3), pp. 380–390. doi: 10.1016/j.stem.2010.07.011.

Škrtić, M. *et al.* (2011) 'Inhibition of Mitochondrial Translation as a Therapeutic Strategy for Human Acute Myeloid Leukemia', *Cancer Cell*. NIH Public Access, 20(5), pp. 674–688. doi: 10.1016/j.ccr.2011.10.015.

Smedley, D. *et al.* (2009) 'BioMart - Biological queries made easy', *BMC Genomics*. BioMed

Central, 10(1), p. 22. doi: 10.1186/1471-2164-10-22.

Smith, M. L. *et al.* (2004) ‘ Mutation of CEBPA in Familial Acute Myeloid Leukemia ’, *New England Journal of Medicine*. Massachusetts Medical Society, 351(23), pp. 2403–2407. doi: 10.1056/nejmoa041331.

Smits, P. *et al.* (2011) ‘Mutation in mitochondrial ribosomal protein MRPS22 leads to Cornelia de Lange-like phenotype, brain abnormalities and hypertrophic cardiomyopathy’, *European Journal of Human Genetics*. Eur J Hum Genet, 19(4), pp. 394–399. doi: 10.1038/ejhg.2010.214.

Snuderl, M. *et al.* (2011) ‘Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma’, *Cancer Cell*. Cancer Cell, 20(6), pp. 810–817. doi: 10.1016/j.ccr.2011.11.005.

Somervaille, T. C. P. and Cleary, M. L. (2006) ‘Identification and characterization of leukemia stem cells in murine MLL-AF9 acute myeloid leukemia’, *Cancer Cell*. Cell Press, 10(4), pp. 257–268. doi: 10.1016/j.ccr.2006.08.020.

Sotgia, F. *et al.* (2012) ‘Mitochondria “fuel” breast cancer metabolism: Fifteen markers of mitochondrial biogenesis label epithelial cancer cells, but are excluded from adjacent stromal cells’, *Cell Cycle*. Taylor and Francis Inc., 11(23), pp. 4390–4401. doi: 10.4161/cc.22777.

Sottoriva, A. *et al.* (2013) ‘Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 110(10), pp. 4009–4014. doi: 10.1073/pnas.1219747110.

Spencer, S. L. *et al.* (2009) ‘Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis’, *Nature*. Nature Publishing Group, 459(7245), pp. 428–432. doi: 10.1038/nature08012.

Ståhl, P. L. *et al.* (2016) ‘Visualization and analysis of gene expression in tissue sections by spatial transcriptomics’, *Science*. American Association for the Advancement of Science, pp. 78–82. doi: 10.1126/science.aaf2403.

Stavropoulou, V. *et al.* (2016) ‘MLL-AF9 Expression in Hematopoietic Stem Cells Drives a

Highly Invasive AML Expressing EMT-Related Genes Linked to Poor Outcome', *Cancer Cell*. Cell Press, 30(1), pp. 43–58. doi: 10.1016/j.ccell.2016.05.011.

Stephens, P. J. *et al.* (2011) 'Massive genomic rearrangement acquired in a single catastrophic event during cancer development', *Cell*. Elsevier, 144(1), pp. 27–40. doi: 10.1016/j.cell.2010.11.055.

Stuart, T. *et al.* (2019) 'Comprehensive Integration of Single-Cell Data', *Cell*. Cell Press, 177(7), pp. 1888-1902.e21. doi: 10.1016/j.cell.2019.05.031.

Subramanian, A. *et al.* (2005) 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(43), pp. 15545–15550. doi: 10.1073/pnas.0506580102.

Süel, G. M. *et al.* (2006) 'An excitable gene regulatory circuit induces transient cellular differentiation', *Nature*. Nature Publishing Group, 440(7083), pp. 545–550. doi: 10.1038/nature04588.

Suganuma, K. *et al.* (2010) 'Energy metabolism of leukemia cells: Glycolysis versus oxidative phosphorylation', *Leukemia and Lymphoma*. Taylor & Francis, 51(11), pp. 2112–2119. doi: 10.3109/10428194.2010.512966.

Sulima, S. O. *et al.* (2017) 'How ribosomes translate cancer', *Cancer Discovery*. American Association for Cancer Research Inc., pp. 1069–1087. doi: 10.1158/2159-8290.CD-17-0550.

Sun, C., Chang, L. and Zhu, X. (2017) 'Pathogenesis of ETV6/RUNX1-positive childhood acute lymphoblastic leukemia and mechanisms underlying its relapse', *Oncotarget*. Impact Journals LLC, pp. 35445–35459. doi: 10.18632/oncotarget.16367.

Sun, Q. Y. *et al.* (2017) 'Ordering of mutations in acute myeloid leukemia with partial tandem duplication of MLL (MLL-PTD)', *Leukemia*. Nature Publishing Group, 31(1), pp. 1–10. doi: 10.1038/leu.2016.160.

Sun, W. and Downing, J. R. (2004) 'Haploinsufficiency of AML1 results in a decrease in the number of LTR-HSCs while simultaneously inducing an increase in more mature progenitors',

Blood. Blood, 104(12), pp. 3565–3572. doi: 10.1182/blood-2003-12-4349.

Svensson, V., Vento-Tormo, R. and Teichmann, S. A. (2018) 'Exponential scaling of single-cell RNA-seq in the past decade', *Nature Protocols*. Nature Publishing Group, pp. 599–604. doi: 10.1038/nprot.2017.149.

Swain, P. S., Elowitz, M. B. and Siggia, E. D. (2002) 'Intrinsic and extrinsic contributions to stochasticity in gene expression', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 99(20), pp. 12795–12800. doi: 10.1073/pnas.162041399.

Szendro, I. G. *et al.* (2013) 'Predictability of evolution depends nonmonotonically on population size', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 110(2), pp. 571–576. doi: 10.1073/pnas.1213613110.

T North, N. A. S. (1999) 'Cbfa2 is required for the formation of intra-aortic hematopoietic clusters', *Development*, 126(11), pp. 2563–75.

Tang, F. *et al.* (2009) 'mRNA-Seq whole-transcriptome analysis of a single cell', *Nature Methods*. Nat Methods, 6(5), pp. 377–382. doi: 10.1038/nmeth.1315.

Tantale, K. *et al.* (2016) 'A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting', *Nature Communications*. Nature Publishing Group, 7. doi: 10.1038/ncomms12248.

TCGA (2013a) 'Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 368(22), pp. 2059–2074. doi: 10.1056/nejmoa1301689.

TCGA (2013b) 'Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 368(22), pp. 2059–2074. doi: 10.1056/nejmoa1301689.

Tehranchi, R. *et al.* (2010) 'Persistent Malignant Stem Cells in del(5q) Myelodysplasia in Remission', *New England Journal of Medicine*. Massachusetts Medical Society, 363(11), pp. 1025–1037. doi: 10.1056/nejmoa0912228.

Teles, J. *et al.* (2013) 'Transcriptional Regulation of Lineage Commitment - A Stochastic Model of Cell Fate Decisions', *PLoS Computational Biology*. PLoS Comput Biol, 9(8). doi: 10.1371/journal.pcbi.1003197.

Thattai, M. and Van Oudenaarden, A. (2001) 'Intrinsic noise in gene regulatory networks', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 98(15), pp. 8614–8619. doi: 10.1073/pnas.151588598.

Thompson, S. L. and Compton, D. A. (2008) 'Examining the link between chromosomal instability and aneuploidy in human cells', *Journal of Cell Biology*. The Rockefeller University Press, 180(4), pp. 665–672. doi: 10.1083/jcb.200712029.

Tighe, J. E. and Calabi, F. (1995) 't(8;21) Breakpoints are Clustered between Alternatively Spliced Exons of MTG8', *Clinical Science*. Portland Press Ltd, 89(3), pp. 215–218. doi: 10.1042/cs0890215.

Tighe, J. E., Daga, A. and Calabi, F. (1995) *Translocation Breakpoints Are Clustered on Both Chromosome 8 and Chromosome 21 in the t(8;21) of Acute Myeloid Leukemia*. Available at: http://ashpublications.org/blood/article-pdf/81/3/592/610145/592.pdf (Accessed: 6 December 2020).

Tomlinson, I. P. M., Novelli, M. R. and Bodmer, W. F. (1996) 'The mutation rate and cancer', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 93(25), pp. 14800–14803. doi: 10.1073/pnas.93.25.14800.

Trapnell, C. *et al.* (2014) 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nature Biotechnology*. Nature Publishing Group, 32(4), pp. 381–386. doi: 10.1038/nbt.2859.

Tschochner, H. and Hurt, E. (2003) 'Pre-ribosomes on the road from the nucleolus to the cytoplasm', *Trends in Cell Biology*. Elsevier Ltd, pp. 255–263. doi: 10.1016/S0962-8924(03)00054-0.

Turner BM (1993) 'Decoding the nucleosome', *Cell*, 75(1), pp. 5–8.

Tzelepis, K. *et al.* (2016) 'A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and

Therapeutic Targets in Acute Myeloid Leukemia', *Cell Reports*. Elsevier B.V., 17(4), pp. 1193–1205. doi: 10.1016/j.celrep.2016.09.079.

Tzoneva, G. *et al.* (2013a) 'Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL', *Nature Medicine*. Nat Med, 19(3), pp. 368–371. doi: 10.1038/nm.3078.

Tzoneva, G. *et al.* (2013b) 'Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL', *Nature Medicine*. Nat Med, 19(3), pp. 368–371. doi: 10.1038/nm.3078.

Urban, E. A. and Johnston, R. J. (2018) 'Buffering and Amplifying Transcriptional Noise During Cell Fate Specification', *Frontiers in Genetics*. Frontiers Media SA, 9. doi: 10.3389/fgene.2018.00591.

Utada, A. S. *et al.* (2005) 'Monodisperse double emulsions generated from a microcapillary device', *Science*. Science, 308(5721), pp. 537–541. doi: 10.1126/science.1109164.

Vafai, S. B. and Mootha, V. K. (2012) 'Mitochondrial disorders as windows into an ancient organelle', *Nature*. Nature, pp. 374–383. doi: 10.1038/nature11707.

Vakana, E. *et al.* (2011) 'Antileukemic effects of AMPK activators on BCR-ABL-expressing cells', *Blood*. Blood, 118(24), pp. 6399–6402. doi: 10.1182/blood-2011-01-332783.

Vincent D Blondel *et al.* (2008) 'Fast unfolding of communities in large networks', *J. Stat. Mech*, p. 10008. doi: 10.1088/1742-5468/2008/10/P10008.

Vlachos, A. *et al.* (2012a) 'Incidence of neoplasia in Diamond Blackfan anemia: A report from the Diamond Blackfan anemia registry', *Blood*. Blood, 119(16), pp. 3815–3819. doi: 10.1182/blood-2011-08-375972.

Vlachos, A. *et al.* (2012b) 'Incidence of neoplasia in Diamond Blackfan anemia: A report from the Diamond Blackfan anemia registry', *Blood*. Blood, 119(16), pp. 3815–3819. doi: 10.1182/blood-2011-08-375972.

Walter, K. *et al.* (2010) 'Aberrant expression of CD19 in AML with t(8;21) involves a poised chromatin structure and PAX5', *Oncogene*. Oncogene, 29(20), pp. 2927–2937. doi:

10.1038/onc.2010.56.

Wang, J. *et al.* (2011) 'Quantifying the Waddington landscape and biological paths for development and differentiation', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(20), pp. 8257–8262. doi: 10.1073/pnas.1017017108.

Wang, L. and Dent, S. Y. R. (2014) 'Functions of SAGA in development and disease', *Epigenomics*. Future Medicine Ltd., pp. 329–339. doi: 10.2217/epi.14.22.

Wang, P. *et al.* (2013) 'Mutations in isocitrate dehydrogenase 1 and 2 occur frequently in intrahepatic cholangiocarcinomas and share hypermethylation targets with glioblastomas', *Oncogene*. NIH Public Access, 32(25), pp. 3091–3100. doi: 10.1038/onc.2012.315.

Wang, Y. L. *et al.* (2008) 'Human ATAC is a GCN5/PCAF-containing acetylase complex with a novel NC2-like histone fold module that interacts with the TATA-binding protein', *Journal of Biological Chemistry*. JBC Papers in Press, 283(49), pp. 33808–33815. doi: 10.1074/jbc.M806936200.

Wang, Y. Y. *et al.* (2005) 'AML1-ETO and C-KIT mutation/overexpression in t(8;21) leukemia: Implication in stepwise leukemogenesis and response to Gleevec', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(4), pp. 1104–1109. doi: 10.1073/pnas.0408831102.

Warner, J. R. (1999) 'The economics of ribosome biosynthesis in yeast', *Trends in Biochemical Sciences*. Trends Biochem Sci, pp. 437–440. doi: 10.1016/S0968-0004(99)01460-7.

Weaver, B. A. A. *et al.* (2007) 'Aneuploidy Acts Both Oncogenically and as a Tumor Suppressor', *Cancer Cell*. Cancer Cell, 11(1), pp. 25–36. doi: 10.1016/j.ccr.2006.12.003.

Weinberger, L. *et al.* (2012a) 'Expression Noise and Acetylation Profiles Distinguish HDAC Functions', *Molecular Cell*. Elsevier, 47(2), pp. 193–202. doi: 10.1016/j.molcel.2012.05.008.

Weinberger, L. *et al.* (2012b) 'Expression Noise and Acetylation Profiles Distinguish HDAC Functions', *Molecular Cell*. Mol Cell, 47(2), pp. 193–202. doi: 10.1016/j.molcel.2012.05.008.

Weissman, I. (2005) 'Stem cell research: Paths to cancer therapies and regenerative medicine',

Journal of the American Medical Association. JAMA, 294(11), pp. 1359–1366. doi: 10.1001/jama.294.11.1359.

Weissman, I. L. and Shizuru, J. A. (2008) 'The origins of the identification and isolation of hematopoietic stem cells, and their capability to induce donor-specific transplantation tolerance and treat autoimmune diseases', *Blood*. American Society of Hematology, 112(9), pp. 3543–3553. doi: 10.1182/blood-2008-08-078220.

Welch, J. S. *et al.* (2012a) 'The origin and evolution of mutations in acute myeloid leukemia', *Cell*. Elsevier, 150(2), pp. 264–278. doi: 10.1016/j.cell.2012.06.023.

Welch, J. S. *et al.* (2012b) 'The origin and evolution of mutations in acute myeloid leukemia', *Cell*. Elsevier, 150(2), pp. 264–278. doi: 10.1016/j.cell.2012.06.023.

Welch, J. S. *et al.* (2012c) 'The origin and evolution of mutations in acute myeloid leukemia', *Cell*. Cell, 150(2), pp. 264–278. doi: 10.1016/j.cell.2012.06.023.

Wernet, M. F. *et al.* (2006) 'Stochastic spineless expression creates the retinal mosaic for colour vision', *Nature*. Nature Publishing Group, 440(7081), pp. 174–180. doi: 10.1038/nature04615.

Whitesides, G. M. (2006) 'The origins and the future of microfluidics', *Nature*. Nature, pp. 368–373. doi: 10.1038/nature05058.

Wiemels, J. L. *et al.* (2002) 'In utero origin of t(8;21) AML1-ETO translocations in childhood acute myeloid leukemia', *Blood*. Blood, 99(10), pp. 3801–3805. doi: 10.1182/blood.V99.10.3801.

Van Wijnen, A. J. *et al.* (2004) 'Nomenclature for Runt-related (RUNX) proteins', *Oncogene*. Oncogene, pp. 4209–4210. doi: 10.1038/sj.onc.1207758.

Williams, D. A. *et al.* (1984) 'Introduction of new genetic material into pluripotent haematopoietic stem cells of the mouse', *Nature*. Nature, 310(5977), pp. 476–480. doi: 10.1038/310476a0.

Williams, M. J. *et al.* (2016) 'Identification of neutral tumor evolution across cancer types', *Nature Genetics*. Nature Publishing Group, 48(3), pp. 238–244. doi: 10.1038/ng.3489.

Williams, M. J., Sottoriva, A. and Graham, T. A. (2019) 'Measuring Clonal Evolution in Cancer with Genomics', *Annual Review of Genomics and Human Genetics*. Annual Reviews Inc., pp. 309–329. doi: 10.1146/annurev-genom-083117-021712.

Wingett, S. W. and Andrews, S. (2018) 'FastQ Screen: A tool for multi-genome mapping and quality control', *F1000Research*. F1000 Research Ltd, 7, p. 1338. doi: 10.12688/f1000research.15931.2.

Wlodarski, M. W. *et al.* (2016) 'Prevalence, clinical characteristics, and prognosis of GATA2-related myelodysplastic syndromes in children and adolescents', *Blood*. American Society of Hematology, 127(11), pp. 1387–1397. doi: 10.1182/blood-2015-09-669937.

Wolf, F. A. *et al.* (2019) 'PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells', *Genome Biology*. BioMed Central Ltd., 20(1), pp. 1–9. doi: 10.1186/s13059-019-1663-x.

Wood, L. D. *et al.* (2007) 'The genomic landscapes of human breast and colorectal cancers', *Science*. Science, 318(5853), pp. 1108–1113. doi: 10.1126/science.1145720.

Wray, J. P. *et al.* (2020) 'Cell cycle corruption in a preleukemic ETV6-RUNX1 model exposes RUNX1 addiction as a therapeutic target in acute lymphoblastic leukemia', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2020.12.22.423823. doi: 10.1101/2020.12.22.423823.

Wu, J. *et al.* (2020) 'A single-cell survey of cellular hierarchy in acute myeloid leukemia', *Journal of Hematology and Oncology*. BioMed Central Ltd, 13(1), p. 128. doi: 10.1186/s13045-020-00941-y.

Wu, X. *et al.* (2012) 'Clonal selection drives genetic divergence of metastatic medulloblastoma', *Nature*. Nature Publishing Group, 482(7386), pp. 529–533. doi: 10.1038/nature10825.

Xie, M. *et al.* (2014) 'Age-related mutations associated with clonal hematopoietic expansion and malignancies', *Nature Medicine*. Nature Publishing Group, 20(12), pp. 1472–1478. doi: 10.1038/nm.3733.

Xu, E. Y., Zawadzki, K. A. and Broach, J. R. (2006) 'Single-Cell Observations Reveal

Intermediate Transcriptional Silencing States', *Molecular Cell*. Mol Cell, 23(2), pp. 219–229. doi: 10.1016/j.molcel.2006.05.035.

Xu, W. *et al.* (2000a) 'Loss of Gcn512 leads to increased apoptosis and mesodermal defects during mouse development', *Nature Genetics*. Nat Genet, 26(2), pp. 229–232. doi: 10.1038/79973.

Xu, W. *et al.* (2000b) 'Loss of Gcn512 leads to increased apoptosis and mesodermal defects during mouse development', *Nature Genetics*. Nature Publishing Group, 26(2), pp. 229–232. doi: 10.1038/79973.

Xu, W. *et al.* (2011) 'Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases', *Cancer Cell*. NIH Public Access, 19(1), pp. 17–30. doi: 10.1016/j.ccr.2010.12.014.

Yamauchi, T. *et al.* (2000) 'Distinct but overlapping roles of histone acetylase PCAF and of the closely related PCAF-B/GCN5 in mouse embryogenesis', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 97(21), pp. 11303–11306. doi: 10.1073/pnas.97.21.11303.

Yan, M. *et al.* (2004) 'Deletion of an AML1-ETO C-terminal NcoR/SMRT-interacting region strongly induces leukemia development', *Proceedings of the National Academy of Sciences of the United States of America*, 101(49), pp. 17186 LP – 17191. doi: 10.1073/pnas.0406702101.

Yan, M. *et al.* (2006) 'A previously unidentified alternatively spliced isoform of t(8;21) transcript promotes leukemogenesis', *Nature Medicine*. Nature Publishing Group, 12(8), pp. 945–949. doi: 10.1038/nm1443.

Yan, M. *et al.* (2009) 'RUNX1/AML1 DNA-binding domain and ETO/MTG8 NHR2-dimerization domain are critical to AML1-ETO9a leukemogenesis', *Blood*. American Society of Hematology, 113(4), pp. 883–886. doi: 10.1182/blood-2008-04-153742.

Yang, D. *et al.* (2011) 'Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer', *JAMA - Journal of the American Medical Association*. JAMA, 306(14), pp. 1557–1565. doi: 10.1001/jama.2011.1456.

Yang, H. *et al.* (2012a) 'IDH1 and IDH2 mutations in tumorigenesis: Mechanistic insights and clinical perspectives', *Clinical Cancer Research*. NIH Public Access, pp. 5562–5571. doi: 10.1158/1078-0432.CCR-12-1773.

Yang, H. *et al.* (2012b) 'IDH1 and IDH2 mutations in tumorigenesis: Mechanistic insights and clinical perspectives', *Clinical Cancer Research*. American Association for Cancer Research, pp. 5562–5571. doi: 10.1158/1078-0432.CCR-12-1773.

Yap, T. A. *et al.* (2012) 'Intratumor heterogeneity: Seeing the wood for the trees', *Science Translational Medicine*. Sci Transl Med. doi: 10.1126/scitranslmed.3003854.

Yates, L. R. *et al.* (2015) 'Subclonal diversification of primary breast cancer revealed by multiregion sequencing', *Nature Medicine*. Nature Publishing Group, 21(7), pp. 751–759. doi: 10.1038/nm.3886.

Yates, L. R. and Campbell, P. J. (2012) 'Evolution of the cancer genome', *Nature Reviews Genetics*. Europe PMC Funders, pp. 795–806. doi: 10.1038/nrg3317.

Yin, Y. W. *et al.* (2015) 'The histone acetyltransferase GCN5 expression is elevated and regulated by c-Myc and E2F1 transcription factors in human colon cancer', *Gene Expression*. Cognizant Communication Corporation, 16(4), pp. 187–196. doi: 10.3727/105221615X14399878166230.

Yu, B. D. *et al.* (1995) 'Altered Hox expression and segmental identity in Mll-mutant mice', *Nature*. Nature, 378(6556), pp. 505–508. doi: 10.1038/378505a0.

Yu, B. D. *et al.* (1998) 'MLL, a mammalian trithorax-group gene, functions as a transcriptional maintenance factor in morphogenesis', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 95(18), pp. 10632–10636. doi: 10.1073/pnas.95.18.10632.

Zelent, A., Greaves, M. and Enver, T. (2004) 'Role of the TEL-AML1 fusion gene in the molecular pathogenesis of childhood acute lymphoblastic leukaemia', *Oncogene*. Oncogene, pp. 4275–4283. doi: 10.1038/sj.onc.1207672.

Zhang, Y. *et al.* (2002) 'Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with

topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8;21) leukemia', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 99(5), pp. 3070–3075. doi: 10.1073/pnas.042702899.

Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biology*. BioMed Central, 9(9), p. R137. doi: 10.1186/gb-2008-9-9-r137.

Zhang, Y. *et al.* (2019) 'The ZZ domain as a new epigenetic reader and a degradation signal sensor', *Critical Reviews in Biochemistry and Molecular Biology*. Taylor and Francis Ltd, pp. 1–10. doi: 10.1080/10409238.2018.1564730.

Zhao, H. *et al.* (2015) 'PARP1- and CTCF-Mediated Interactions between Active and Repressed Chromatin at the Lamina Promote Oscillating Transcription', *Molecular Cell*. Cell Press, 59(6), pp. 984–997. doi: 10.1016/j.molcel.2015.07.019.

Zhao, S. *et al.* (2009) 'Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1α', *Science*, 324(5924), pp. 261–265. doi: 10.1126/science.1170944.

Zhou, W. *et al.* (2019) 'Global prediction of chromatin accessibility using small-cell-number and single-cell RNA-seq', *Nucleic acids research*. NLM (Medline), 47(19), p. e121. doi: 10.1093/nar/gkz716.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., *et al.* (2017) 'Comparative Analysis of Single-Cell RNA Sequencing Methods', *Molecular Cell*. Cell Press, 65(4), pp. 631-643.e4. doi: 10.1016/j.molcel.2017.01.023.

Ziegenhain, C., Vieth, B., Parekh, S. and Heyn, H. (2017) 'Comparative Analysis of Single-Cell RNA Sequencing Methods', *Molecular Cell*, 65, pp. 631-643.e4. doi: 10.1016/j.molcel.2017.01.023.

Zoller, B. *et al.* (2015) 'Structure of silent transcription intervals and noise characteristics of mammalian genes', *Molecular Systems Biology*. EMBO, 11(7), p. 823. doi: 10.15252/msb.20156257.

Zong, C. *et al.* (2012) 'Genome-wide detection of single-nucleotide and copy-number variations of a single human cell', *Science*. American Association for the Advancement of

Science, 338(6114), pp. 1622–1626. doi: 10.1126/science.1229164.

Zuber, J. *et al.* (2009) ‘Mouse models of human AML accurately predict chemotherapy response’, *Genes and Development*. Cold Spring Harbor Laboratory Press, 23(7), pp. 877–889. doi: 10.1101/gad.1771409.

Annexure-A: Analysis Scripts

A.1 Single-cell RNA sequencing analysis

Pre-processing and filtering

```
# library loading

library(Seurat)
library(Matrix)
library(mclust)
library(dplyr)
library(reshape)
library(rgl)
library(pca3d)
library(ggplot2)

## creating a Seurat object

amlseu_500_3 <- CreateSeuratObject(raw.data = amlseu_500_3.data,
min.cells = 53, min.genes = 500, project = "AML1_ETO")

# reading a Seurat object

readRDS("~/Desktop/mm10_2/amlseu_500_3.rds")

amlseu_500_3 <-
readRDS("~/Desktop/mm10_2/amlseu_500_3_clusterid.rds")

## calculating percentage mitochondrial content

mito.genes <- grep(pattern = "^MT-", x = rownames(x =
amlseu_500_3@data), value = TRUE)
percent.mito <- Matrix::colSums(amlseu_500_3@raw.data[mito.genes,
])/Matrix::colSums(amlseu_500_3@raw.data)

## adding gene counts, UMI counts and percentage mitochondrial
content as metadata

pbmc <- AddMetaData(object = amlseu_500_3, metadata = percent.mito,
col.name = "percent.mito")
```

```
VlnPlot(object = amlseu_500_3, features.plot = c("nGene", "nUMI",
"percent.mito"), nCol = 3)

## plotting relationship between UMI counts and gene counts or
mitochondrial content

GenePlot(object = amlseu_500_3, gene1 = "nUMI", gene2 =
"percent.mito")
GenePlot(object = amlseu_500_3, gene1 = "nUMI", gene2 = "nGene")

## Filtering cells with low and high threshold

amlseu_500_3 <- FilterCells(object = amlseu_500_3, subset.names =
c("nGene", "percent.mito"), low.thresholds = c(200, -Inf),
high.thresholds = c(2500, 0.05))
```

Detection of variable genes

```
amlseu_500_3 <- FindVariableGenes(object = amlseu_500_3,
mean.function = ExpMean, dispersion.function = LogVMR, x.low.cutoff
= 0.0125, x.high.cutoff = 3, y.cutoff = 0.5)


length(x = amlseu_500_3@var.genes) ## getting the number of variable
genes
```

Differential expression calculation using DESeq2

```
amlseu_500_3 <- SetAllIdent(object = amlseu_500_3, id = 'ident')

diff_56_46_deseq <- FindMarkers(object = amlseu_500_3, ident.1 =
'IN56', ident.2 = 'IN46', test.use = "DESeq2")

amlseu_500_3 <- ScaleData(object = amlseu_500_3, vars.to.regress =
c("nUMI", "percent.mito")). ## scaling the data
```

Linear Dimensionality reduction analysis

```
amlseu_500_3 <- RunPCA(object = amlseu_500_3, pc.genes =
amlseu_500_3@var.genes, do.print = TRUE, pcs.print = 1:5,
genes.print = 5)
```

```
PCAPlot(object = amlseu_500_3, dim.1 = 1, dim.2 = 2) ## plotting PCA

PCHeatmap(object = amlseu_500_3, pc.use = 10, cells.use = 500,
do.balanced = TRUE, label.columns = FALSE)

amlseu_500_3 <- JackStraw(object = amlseu_500_3, num.replicate =
100, display.progress = FALSE) ## Jackstraw plot for estimating the
significant PCs

JackStrawPlot(object = amlseu_500_3, PCs = 1:12) ## visualizing
Jackstraw plot

PCElbowPlot(object = amlseu_500_3) ## elbow plot to determine the
significant PCs
```

Graph based clustering analysis

```
amlseu_500_3 <- RunPCA(object = amlseu_500_3, pc.genes =
amlseu_500_3@var.genes, do.print = TRUE, pcs.print = 1:5,
genes.print = 5)

amlseu_500_3 <- FindClusters(object = amlseu_500_3, reduction.type =
"pca", dims.use = 1:10, resolution = 0.6, print.output = 0, save.SNN
= TRUE)

PrintFindClustersParams(object = amlseu_500_3)
```

Non-linear dimensionality reduction analysis

```
amlseu_500_3 <- RunTSNE(object = amlseu_500_3, dims.use = 1:10,
do.fast = TRUE)
TSNEPlot(object = amlseu_500_3)
```

Creation of Cell Data Set object for pseudotime analysis using Monocle v3.0

```
## reading barcodes from combined gene expression matrix

amlseu_sample_sheet <-
read.table("~/Desktop/mm10_2/total.barcodes_monocle.txt", sep =
"\t")
```

```
## reading gene IDs from combined gene expression matrix

amlseu_gene_annotation_2 <-
read.table("~/Desktop/mm10_2/total.genes_monocle.txt", sep = "\t")

rownames(amlseu_sample_sheet) = amlseu_sample_sheet$V2

colnames(amlseu_gene_annotation_2) <- c("V1", "gene_short_name")

rownames(amlseu_gene_annotation_2) =
amlseu_gene_annotation_2$gene_short_name

pd_2 <- new("AnnotatedDataFrame", data = amlseu_sample_sheet)
fd_2 <- new("AnnotatedDataFrame", data = amlseu_gene_annotation_2)

amlseu_500_3_mono.data <- Read10X(data.dir = "~/Desktop/mm10_2/")

## obtaining raw expression matrix using Seurat object

amlseu_mono <- CreateSeuratObject(raw.data = amlseu_500_3_mono.data,
project = "AMLETO_Monocle", min.cells = 53, min.genes = 500)
expression_matrix <- amlseu_mono@data

## creating a new cell data set object

aml_m3_cds <- new_cell_data_set(expression_matrix, cell_metadata =
amlseu_sample_sheet, gene_metadata = amlseu_gene_annotation_2)

## performing pre-processing and normalization

aml_m3_cds = preprocess_cds(aml_m3_cds, num_dim = 100)
plot_pc_variance_explained(aml_m3_cds)

## clustering the cells using k-means clustering

aml_m3_cds <- cluster_cells(aml_m3_cds)
plot_cells(aml_m3_cds)

## colouring the cells based on sample ID
```

```
colo <-
c(rep("IN42",515),rep("IN44",497),rep("IN46",374),rep("IN56",353))

colData(aml_m3_cds)$sampleID <- colo

plot_cells(aml_m3_cds, color_cells_by = 'sampleID')
```

Performing dimensionality reduction analysis

```
aml_m3_cds <- reduce_dimension(aml_m3_cds, reduction_method =
"tSNE")

## clustering the cells using k-means clustering and overlaying on
tSNE plot

aml_m3_cds <- cluster_cells(aml_m3_cds, reduction_method = "tSNE")
```

Pseudotime trajectory analysis

```
## learning the trajectory

aml_m3_cds <- learn_graph(aml_m3_cds)

plot_cells(aml_m3_cds, color_cells_by = 'sampleID',
label_cell_groups = FALSE, label_leaves = TRUE, label_branch_points
= TRUE, graph_label_size = 1.5)

## ordering the cells in pseudotime

aml_m3_cds = order_cells(aml_m3_cds)

plot_cells(aml_m3_cds, color_cells_by = "pseudotime",
label_cell_groups = 'sampleID', label_leaves = FALSE,
label_branch_points = FALSE, graph_label_size = 1.5)

## obtaining the principal node

get_earliest_principal_node <- function(aml_m3_cds, id = "IN46"){
+ cell_ids <- which(colData(aml_m3_cds)[,"sampleID"] == id)
```

```
+ closest_vertex <-
aml_m3_cds@principal_graph_aux[["UMAP"]]$pr_graph_cell_proj_closest_
vertex
+ closest_vertex <- as.matrix(closest_vertex[colnames(aml_m3_cds),
])
+ root_pr_nodes <-
igraph::V(principal_graph(aml_m3_cds)[["UMAP"]])$name[as.numeric(nam
es(which.max(table(closest_vertex[cell_ids, ]))))]
+ root_pr_nodes}

## ordering the cells relative to principal node

aml_m3_cds = order_cells(aml_m3_cds, root_pr_nodes =
get_earliest_principal_node(aml_m3_cds))
plot_cells(aml_m3_cds, color_cells_by = "pseudotime",
label_cell_groups = FALSE, label_leaves = FALSE, label_branch_points
= FALSE, graph_label_size = 1.5)
```

Pairwise distance calculation

```
# Subsetting required matrix of cells

Non_LeuP_amlseu <- SubsetData(object = amlseu_500_3, ident.use =
c('B-cells', 'LMPPs', 'CMPs', 'Monocytes'))

# Calculate mean, CV and DM

means_Non_LeuP <- rowMeans(as.matrix(Non_LeuP_amlseu@data))
cv2_Non_LeuP <- apply(as.matrix(Non_LeuP_amlseu@data), 1,
var)/means_Non_LeuP^2
dm_Non_LeuP <- DM(means_Non_LeuP, cv2_Non_LeuP)

# Sort based on DM to identify top 500 highly variable genes

dm_Non_LeuP_sorted <- sort(dm_Non_LeuP, decreasing = TRUE)
top_500_Non_LeuP <- dm_Non_LeuP_sorted[1:500]

# extracting matrix corresponding to highly variable genes
Non_LeuP_subset_matrix <-
Non_LeuP_amlseu@data[names(top_500_Non_LeuP),]
dim(Non_LeuP_subset_matrix)
```

```
# calculating spearman correlation
Non_LeuP_spearman <- cor(x = as.matrix(Non_LeuP_subset_matrix), y =
NULL, method = "spearman")

# calculating distance, d
d_Non_LeuP <- sqrt((1-Non_LeuP_spearman)/2)
dim(d_Non_LeuP)

# for plotting merging every cell type 'd' in single vector
d_Non_LeuP_vector <- as.vector(d_Non_LeuP)
d_NonLeuP_w_LeuP <- c(d_Non_LeuP_vector, d_LeuP_vector)
sampid_NonLeuP_w_LeuP <- c(rep("Non-Leukaemic Progenitors",
length(d_Non_LeuP_vector)), rep("Leukaemic Progenitors",
length(d_LeuP_vector)))

# making a dataframe for 'd' and 'sampleID'
total_NonLeuP_w_LeuP_dataframe <- data.frame(d_NonLeuP_w_LeuP,
sampid_NonLeuP_w_LeuP)
dim(total_NonLeuP_w_LeuP_dataframe)

# assigning column names for defining aesthetics in ggplot
colnames(total_NonLeuP_w_LeuP_dataframe) <- c("Pairwise_Distance",
"Cell_Type")
p_NonLeuP_w_LeuP_total <- ggplot(total_NonLeuP_w_LeuP_dataframe,
aes(factor(Cell_Type), Pairwise_Distance))
p_NonLeuP_w_LeuP_total + geom_violin(aes(fill = factor(Cell_Type)))
# colours based on cell type
```

A.2 Single-cell ATAC sequencing analysis

Filtering the lower quality cells

```
## loading single-cell ATAC sequencing individual peak vs gene
matrix for control DMSO and treated MB-3 cells

dmso_mb3_file_name="/home/sg823/scATAC_Kasumi_analysis3_01092018/dms
o_mb3_raw_peaks_sorted_merged.txt"

import numpy as np
dmso_mb3_column_range=range(3,91)
print dmso_mb3_column_range
```

```
dmso_mb3_data=np.loadtxt(dmso_mb3_file_name,delimiter='\t',dtype=np.
int,usecols=dmso_mb3_column_range)

print "dmso_mb3_matrix"
print dmso_mb3_data
print dmso_mb3_data.shape

## to obtain peak information, chromosome start and stop location
for each peak in input file

dmso_mb3_peak_info=[]
with open(dmso_mb3_file_name,'r') as f:

      for line in f:
            peak_info=line.split('\t')[0:3]
            dmso_mb3_peak_info.append(peak_info)

## to count number of 1's for each cell and impose a threshold of
15% 1's which is 13 in this case (15% of 88)

dmso_mb3_binary=(dmso_mb3_data>0)
dmso_mb3_sum_row= np.sum(dmso_mb3_binary,axis=1)
dmso_mb3_indices= (dmso_mb3_sum_row>=13)

## dmso_mb3_indices is a boolean array of length number of peaks
## selecting those rows from peak info which have a value of indices
as True

dmso_mb3_filt_peak_info=[]
print len(dmso_mb3_peak_info)
for i in range(len(dmso_mb3_peak_info)):
      if dmso_mb3_indices[i]:
            dmso_mb3_filt_peak_info.append(dmso_mb3_peak_info[i])

## dmso_mb3_indices is boolean array for obtaining data from
dmso_mb3_data

dmso_mb3_filt_peak_data=dmso_mb3_data[dmso_mb3_indices,:]

## peak_info an array of chromosome location, start, stop
```

```
## peak_data an array of all binary values for each cell and each
peak
## peak_info_string is a string having each value representing
chromosome location, start and stop
## peak_data_string is a string each value of which is 88 binary
values for each peak

dmso_mb3_filt_peak_data_file='/home/sg823/scATAC_Kasumi_analysis3_01
092018/dmso_mb3_raw_15_filt_peaks.txt'
with open(dmso_mb3_filt_peak_data_file,'w') as f:
      for i in range(len(dmso_mb3_filt_peak_info)):
            peak_info=dmso_mb3_filt_peak_info[i]
            peak_data=dmso_mb3_filt_peak_data[i,:]
            peak_info_string='\t'.join(peak_info)
            peak_data_string='\t'.join([str(val) for val in
peak_data])
            f.write('\t'.join([peak_info_string,peak_data_string]))
            f.write('\n')
```

Jaccard distance calculation

```
## loading filtered single cell ATAC sequencing matrix having peak
vs genes for all cells passing filtration criteria

dmso_mb3_file_name="/home/sg823/scATAC_Kasumi_analysis3_01092018/dms
o_mb3_common_all_peaks_infogain_w_chrm_binary_0_1_2_labels.txt"
import numpy as np
dmso_mb3_column_range=range(10,98)
print dmso_mb3_column_range
dmso_mb3_data=np.loadtxt(dmso_mb3_file_name,delimiter='\t',usecols=d
mso_mb3_column_range)

print "dmso_mb3_matrix"
print dmso_mb3_data
print dmso_mb3_data.shape
dmso_mb3_mean=np.mean(dmso_mb3_data,axis=1)
print "dmso_mb3_mean"
print dmso_mb3_mean
print dmso_mb3_data[:,0]
from scipy.spatial import distance
```

```
dmso_mb3_columns=dmso_mb3_data.shape[1]
dmso_mb3_rows=dmso_mb3_data.shape[0]

## padding zeroes in final matrix

dmso_mb3_jaccard_distances=np.zeros((dmso_mb3_columns,dmso_mb3_colum
ns))

## calculation of jaccard distance for individual cell with respect
to each cell

for i in range(dmso_mb3_columns):
      for j in range(dmso_mb3_columns):

      dmso_mb3_jaccard_distances[i,j]=distance.jaccard(dmso_mb3_data
[:,i],dmso_mb3_data[:,j])
print dmso_mb3_jaccard_distances
np.savetxt('/home/sg823/scATAC_Kasumi_analysis3_01092018/dmso_mb3_co
mmon_all_cells_jaccard_distances.txt',dmso_mb3_jaccard_distances,del
imiter='\t')
```

Dimensionality reduction analysis using tSNE

```
## loading filtered single cell ATAC sequencing matrix having peak
vs genes for all cells passing filtration criteria

dmso_mb3_comb_file_name="/home/sg823/scATAC_Kasumi_analysis3_0109201
8/dmso_mb3_comb_filt_peaks.txt"

import numpy as np
np.random.seed(50)      ## start from same seed to avoid variability
in tsne
dmso_mb3_column_range=range(3,91)
print dmso_mb3_column_range
dmso_mb3_data=np.loadtxt(dmso_mb3_comb_file_name,delimiter='\t',usec
ols=dmso_mb3_column_range)

## creating a transpose of dmso_mb3_data matrix

dmso_mb3_data_t=dmso_mb3_data.T
print dmso_mb3_data_t
```

```
print dmso_mb3_data_t.shape
from sklearn.manifold import TSNE
from scipy.spatial import distance
dmso_mb3_data_t_embedded =
TSNE(n_components=2,metric=distance.jaccard).fit_transform(dmso_mb3_
data_t)
print dmso_mb3_data_t_embedded
print dmso_mb3_data_t_embedded.shape
import matplotlib
matplotlib.use('agg')
import matplotlib.pyplot as plt

## creating a list of colours which contains 1 for all DMSO cells

dmso_mb3_colors=[1 for i in range(0,38)]

## creating a list of colours which contains 2 for all mb3 cells

mb3_colors=[2 for i in range(0,50)]
print dmso_mb3_colors
print mb3_colors

## creating an array of 88 elements with initial 38 values as 1 and
next 50 as 2, for colouring cells from DMSO and MB-3 differently

dmso_mb3_colors.extend(mb3_colors)

print dmso_mb3_colors

## plots DMSO and MB-3 cells on tSNE without any colour code

plt.figure()
plt.scatter(dmso_mb3_data_t_embedded[:,0],
dmso_mb3_data_t_embedded[:,1])
plt.savefig('/home/sg823/scATAC_Kasumi_analysis3_01092018/dmso_mb3_c
omb_filt_tsne_wo_color.png')

## plots DMSO and MB-3 cells on tSNE with colour based on values in
dmso_mb3_colors array

plt.figure()
```

```
plt.scatter(dmso_mb3_data_t_embedded[:,0],
dmso_mb3_data_t_embedded[:,1], c=dmso_mb3_colors)
plt.legend()
plt.savefig('/home/sg823/scATAC_Kasumi_analysis3_01092018/dmso_mb3_c
omb_filt_tsne_coloured_groups.png')
```

k-medoid clustering

```
from Pycluster import kmedoids
import numpy as np
dmso_mb3_jaccard_file_name='/home/sg823/scATAC_Kasumi_analysis3_0109
2018/dmso_mb3_common_all_cells_jaccard_distances.txt'
dmso_mb3_jaccard_data=np.loadtxt(dmso_mb3_jaccard_file_name,
delimiter='\t')

print dmso_mb3_jaccard_data.shape
for n in range(2,11):

clusterid,error,nfound=kmedoids(dmso_mb3_jaccard_data,nclusters=n,np
ass=20)
        print n. ## print the number of clusters
        for val in clusterid:
            print val,',',
        print '\n'
        print error
        print nfound
        print '\n\n'
```

Projection of clusters on tSNE plot

```
dmso_mb3_comb_file_name="/home/sg823/scATAC_Kasumi_analysis3_0109201
8/dmso_mb3_common_all_peaks_infogain_w_chrm_binary_0_1_2_labels.txt"

import numpy as np
np.random.seed(50)
dmso_mb3_column_range=range(10,98)
print dmso_mb3_column_range
dmso_mb3_data=np.loadtxt(dmso_mb3_comb_file_name,delimiter='\t',usec
ols=dmso_mb3_column_range)

## generating transpose of dmso_mb3_data matrix
```

```
dmso_mb3_data_t=dmso_mb3_data.T
print dmso_mb3_data_t
print dmso_mb3_data_t.shape
from sklearn.manifold import TSNE
from scipy.spatial import distance
dmso_mb3_data_t_embedded=TSNE(n_components=2,metric=distance.jaccard
).fit_transform(dmso_mb3_data_t). ## obtaining the embeddings for
tSNE
print dmso_mb3_data_t_embedded
print dmso_mb3_data_t_embedded.shape
import matplotlib
matplotlib.use('agg')
import matplotlib.pyplot as plt
dmso_mb3_clusters=[21 , 21 , 21 , 21 , 53 , 21 , 21 , 21 , 21 , 21 ,
21 , 45 , 21 , 21 , 21 , 21 , 21 , 21 , 21 , 21 , 21 , 21 , 21
, 21 , 45 , 21 , 21 , 21 , 21 , 21 , 21 , 21 , 21 , 21 , 21 ,
21 , 45 , 45 , 45 , 45 , 45 , 45 , 45 , 45 , 53 , 45 , 53 , 53
, 45 , 53 , 53 , 53 , 45 , 45 , 45 , 53 , 45 , 53 , 53 , 53 , 21 ,
45 , 45 , 45 , 45 , 45 , 21 , 45 , 45 , 45 , 45 , 45 , 21 , 45 , 53
, 53 , 53 , 45 , 45 , 45 , 45 , 45 , 45 , 45 , 21 , ]

dmso_mb3_clusters=[1 for i in range(0,38)] + [2 for i in
range(0,50)]
plt.figure()
plt.scatter(dmso_mb3_data_t_embedded[:,0],
dmso_mb3_data_t_embedded[:,1], c=dmso_mb3_clusters)
plt.savefig('/home/sg823/scATAC_Kasumi_analysis3_01092018/dmso_mb3_c
ommon_kmed_cells_3_clusters.png')
```

Calculation of differential accessibility peaks

```
dmso_mb3_file_name="/home/sg823/scATAC_Kasumi_analysis3_01092018/dms
o_mb3_raw_15_filt_peaks.txt"
import numpy as np
import statsmodels.stats.multitest
import scipy.stats

dmso_mb3_column_range=range(3,91)
dmso_mb3_infogain_data=np.loadtxt(dmso_mb3_file_name,
delimiter='\t',usecols=dmso_mb3_column_range)
```

```
dmso_mb3_row_range=range(0,dmso_mb3_infogain_data.shape[0])
dmso_mb3_infogain_data_col_range=range(0,88)
pvalue_array=[]
data_array=[]
for i in dmso_mb3_row_range:
      count_zero=0
      count_nonzero=0
      count_g1zero=0
      count_g1nonzero=0
      count_g2zero=0
      count_g2nonzero=0
      for j in dmso_mb3_infogain_data_col_range:
            if dmso_mb3_infogain_data[i][j]==0:
                  count_zero+=1
            else:
                  count_nonzero+=1
            if j>37:
                  if dmso_mb3_infogain_data[i][j]==0:
                        count_g2zero+=1
                  else:
                        count_g2nonzero+=1
            if j<=37:
                  if dmso_mb3_infogain_data[i][j]==0:
                        count_g1zero+=1
                  else:
                        count_g1nonzero+=1

      p_zero=1.0*count_zero/(count_zero+count_nonzero)
      p_nonzero=1-p_zero
      pk=[p_zero,p_nonzero]
      entropy=scipy.stats.entropy(pk,base=2)  ## calculating global
entropy

## calculating entropy for DMSO cells

      p_g1zero=1.0*count_g1zero/(count_g1zero+count_g1nonzero)
      p_g1nonzero=1-p_g1zero
      pk_g1=[p_g1zero,p_g1nonzero]
      g1_entropy=scipy.stats.entropy(pk_g1,base=2)

## calculating entropy for MB-3 cells
```

```
      p_g2zero=1.0*count_g2zero/(count_g2zero+count_g2nonzero)
      p_g2nonzero=1-p_g2zero
      pk_g2=[p_g2zero,p_g2nonzero]
      g2_entropy=scipy.stats.entropy(pk_g2,base=2)

      oddsratio,pvalue=scipy.stats.fisher_exact([[count_g1zero,count
_g2zero],[count_g1nonzero,count_g2nonzero]])
      qvalue=pvalue*dmso_mb3_infogain_data.shape[0]
      info_gain=entropy-(38.0/88)*g1_entropy-(50.0/88)*g2_entropy

      celltype=0
      threshold=0
      if p_g2nonzero-p_g1nonzero>threshold:
            celltype=2

      elif p_g1nonzero-p_g2nonzero>threshold:
            celltype=1
      data_array.append([info_gain,pvalue,qvalue,oddsratio,celltype]
)
      pvalue_array.append(pvalue)
rejected,pvalue_bj=statsmodels.stats.multitest.fdrcorrection(pvalue_
array,alpha=0.05,method='indep', is_sorted=False). ## calculation of
adjusted p_value using Benjamin Hochberg method

for i in dmso_mb3_row_range:
      str_data = [str(val) for val in data_array[i]]
      rejected_curr = 1 if rejected[i] else 0
```

Annexure-B

B.1: Definition of haematopoietic compartments for flow cytometry

Cell population	Gating strategy
HSC	Lin^{-} $cKit^{+}$ $Sca1^{+}$ $CD34^{-}$ $CD135^{-}$
MPP	Lin^{-} $cKit^{+}$ $Sca1^{+}$ $CD34^{+}$ $CD135^{-}$
LMPP	Lin^{-} $cKit^{+}$ $Sca1^{+}$ $CD34^{+}$ $CD135^{+}$
CMP	Lin^{-} $cKit^{+}$ $Sca1^{-}$ $CD34^{+/lo}$ $CD16/32^{lo}$
GMP	Lin^{-} $cKit^{+}$ $Sca1^{+}$ $CD34^{+}$ $CD16/32^{hi}$
MEP	Lin^{-} $cKit^{+}$ $Sca1^{+}$ $CD34^{-}$ $CD16/32^{-}$

B.2: Primers for Genotyping

Primer	Forward primer (5'-3')	Reverse primer 1 (5'-3')	Reverse primer 2 (5'-3')
Mx1- Cre	CGTACTGACGGTGGGAGAAT	TGCATGATCTCCGGTATTGA	-
Kat2a	CACAGAGCTTCTTGGAGACC	GGCTTGATTCCTGTACCTCC	-
Idh1	ATAGTCTGGACCATGGGACC	TGTTAGTCCCAACCCCTTCC	GACAAACTGACAGGCTGCAA

B.3: Primers to confirm excision/mutation recombination

Primer	Forward primer (5'-3')	Reverse primer 1 (5'-3')	Reverse primer 2 (5'-3')
Kat2a IN	CAACTTCCCCAAGGTATGGA	CGGGGACCTTAGACTTGTGA	-

Kat2a OUT	AGTCTGGGCTGTTTCCATGT	GCCCGTTGTAGAATGTCTGG	-
Idh1	ATAGTCTGGACCATGGGACC	TGTTAGTCCCAACCCCTTCC	GACAAACTGACAGGCTGCAA

B.4: List of antibodies and fluorescent dyes

Antibody	Fluorochrome	Catalogue ID	Clone	Dilution	Supplier
Annexin V	APC	640919	-		BioLegend
CD45R/ B220	APC-Cy7	103223	RA3-6B2	1:50	BioLegend
CD45R/ B220	PerCP-Cy5.5	103235	RA3-6B2	1:100	BioLegend
CD117/c-Kit	APC-Cy7	105826	2B8	1:50	BioLegend
CD11b/Mac1	AF700	101222	M1/70	1:200	BioLegend
CD14	PE-Cy7	123315	Sa14-2	1:100	BioLegend
CD16/32/FcγR	PE	101308	93	1:100	BioLegend
CD24	BV510	101831	M1/69	1:100	BioLegend
CD34	APC	128612	HM34	1:100	BioLegend
F4/80	PE	123109	BM8	1:100	BioLegend
Gr1	PB	108430	RB6-8C5	1:100	BioLegend
Hoechst 33342	-	H3570	-	1:10000	Invitrogen
Hoechst 33258	-	H3569	-	1:10000	Invitrogen
Sca1	PE-Cy7	108114	D7	1:100	BioLegend
Streptavidin	BV421	405226	-	1:200	BioLegend
Streptavidin	BV605	405229	-	1:200	BioLegend
CD45R/B220 (Lin)	Biotin	103204	RA3-6B2	1:300	BioLegend
Ter119 (Lin)	Biotin	116204	Ter119	1:300	BioLegend
Gr1 (Lin)	Biotin	108404	RB6-8C5	1:300	Biolegend
CD3e (Lin)	Biotin	100304	145–2 C11	1:300	BioLegend
CD11b (Lin)	Biotin	101204	M1/70	1:300	BioLegend
Nanobeads	Streptavidin	76447	-	1:10	BioLegend

Click-iT Cell Reaction Buffer Kit	AF647 azide	A10277	-	1:500	Invitrogen

B.5: List of cell culture reagents

Reagent	Catalogue ID	Supplier
RPMI	R8758	Sigma
IMDM	I3390	Sigma
DMEM	D6429	Sigma
FBS	F9665	Sigma
L-Glutamine	G7513	Sigma
Antimycotic	A5955	Sigma
Trypsin	59418C	Sigma
PBS	D8537	Sigma
Methylcellulose	M3434	Stem Cell Technologies
Mouse IL-3	213-13	Peprotech
Mouse IL-6	500-P56	Peprotech
Mouse SCF	500-P71	Peprotech

B.6: List of Molecular Biology reagents

Reagent	Catalogue ID	Supplier
DNA extraction kit	K1820-01	Invitrogen
SYBR master mix	UF-LSMT-B0701	Takyon
PCR master mix	203643	Qiagen
Agarose	BP160-100	Fisher Bioregents
BSA	BP9700-100	Fisher Scientific
SYBR safe DNA gel stain	NBS-SV	Invitrogen